UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO
DEPARTAMENTO DE GENÉTICA

TAMARA SOLEDAD FRONTANILLA RECALDE

Diversidade genética de populações globais inferida por STRs autossômicos utilizados para identificação humana genotipados a partir de genomas completos

RIBEIRÃO PRETO - SP

2023

TAMARA SOLEDAD FRONTANILLA RECALDE

Diversidade genética de populações globais inferida por STRs autossômicos utilizados para identificação humana genotipados a partir de genomas completos

Tese de Doutorado apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, para obtenção do título de Doutor em Ciências, área de concentração em Genética. Versão corrigida. A versão original encontra-se disponível tanto na Biblioteca da Unidade que aloja o Programa, quanto na Biblioteca Digital de Teses e Dissertações da USP (BDTD).

Orientador: Prof. Dr. Celso Teixeira Mendes Junior

RIBEIRÃO PRETO – SP
2023

Frontanilla TS. Diversidade genética de populações globais inferida por STRs autossômicos utilizados para identificação humana genotipados a partir de genomas completos. Tamara Soledad Frontanilla Recalde; orientador: Prof. Dr. Celso Teixeira Mendes Junior. Ribeirão Preto, 2023;122p.

Tese de Doutorado apresentada à Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo para obtenção do título de Doutor em Ciências, área de concentração Genética – USP/ FMRP/ Departamento de Genética.

Palavras-chave: Microssatélites; Repetições curtas em tandem; sequenciamento massivo em paralelo; bioinformática; genética forense.

Keywords: Microsatellites; Short tandem repeats; massively parallel sequencing; bioinformatics; forensic genetics.

Nome: Frontanilla Recalde Tamara Soledad

Título: Diversidade genética de populações globais inferida por STRs autossômicos utilizados para identificação humana genotipados a partir de genomas completos.

Title: Genetic diversity of worldwide populations inferred by autosomal STRs used for human identification genotyped from whole genome sequencing.

Tese de Doutorado apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, para obtenção do título de Doutor em Ciências, área de concentração em Genética.

Orientador: Prof. Dr. Celso Teixeira Mendes Junior

Aprovado em: 15 de março de 2023

Banca Examinadora:

Presidente: Prof. Dr. Celso Teixeira Mendes Junior
Instituição: Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto - USP
Assinatura: ____

Prof. Dr. Ronaldo Carneiro da Silva Junior
Instituição: Polícia Federal
Julgamento: ____
Assinatura: ____

Prof. Dr. Elizeu Fagundes de Carvalho
Instituição: Universidade do Estado do Rio de Janeiro
Julgamento: ____
Assinatura: ____

Prof. Dra. Andrea Rita Marrero
Instituição: Universidade Federal de Santa Catarina
Julgamento: ____
Assinatura: ____

O presente trabalho foi realizado com apoio do CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brasil.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001

Ao meu marido, pelo apoio incondicional e a ajuda em todos esses anos de estudo.

AGRADECIMENTOS

A Deus, por ter me dado a possibilidade de seguir meus sonhos, e a capacidade para descobri-los.

À minha família, pelo apoio incondicional. Especialmente ao meu marido por ter me apoiado em tudo sempre, por ter confiando em mim, me ensinado com tanta paciência todos os conceitos de bioinformática e por ser um pilar fundamental na minha vida. Também a meus pais, Mirian e Narciso, por me apoiarem, sem dúvidas, em todos os meus sonhos e projetos. E, um carinho especial aos meus tios, Maria Lucia e Álvaro por terem me dado uma casa no Brasil, muito carinho e apoio sempre.

Ao meu amigo e parceiro de pesquisas Guilherme Valle Silva. Com certeza não teria conseguido esses resultados sem o apoio incondicional dele, o ensino e as tardes de trabalho, risadas e pesquisa.

Ao meu orientador, o Prof. Dr. Celso Teixeira Mendes Junior a quem admiro, e respeito muito. Obrigada pela confiança, paciência, oportunidades profissionais, motivação e por contribuir de maneira inigualável na minha formação profissional.

Aos meus colegas e amigos do Departamento; Thássia, Guilherme, Malu, Letícia, Vitor, Alison, Amanda, Nadia, Matheus, Hiago, Luciellen, Mariana, Luiza, por tantas risadas, apoio e por fazerem as tarefas e o trabalho muito mais divertidas.

Com todos e cada um de vocês, para sempre, muito obrigada.

Don't try to build a wall. Don't start a day and say “I am going to build the biggest and most colossal wall that has ever been built. Instead of that say "I am going to lay this brick as perfectly as a brick can be laid" and do that every day, and soon, you'll have a wall".

Will Smith

RESUMO

FRONTANILLA, Tamara Soledad. Diversidade genética de populações globais inferida por STRs autossômicos utilizados para identificação humana genotipados a partir de genomas completos 2023, 122 páginas. [Tese de doutorado]. Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, São Paulo, 2020.

Os marcadores STRs compreendem cerca de 3% do genoma humano. Uma estratégia para compreender melhor a constituição genética de uma população é estimar a distribuição das frequências alélicas de marcadores STRs. A técnica de PCR junto com a eletroforese capilar são, até hoje, as técnicas de escolha para a genotipagem de marcadores STR. As tecnologias de NGS revolucionaram as ciências biológicas, permitindo o sequenciamento simultâneo de muitas amostras de DNA ou RNA em um curto período de tempo. Com os avanços tecnológicos e a utilização cada vez mais frequente das técnicas de NGS surge a necessidade de testar a efetividade dessa técnica na genotipagem de marcadores STRs. Algumas ferramentas foram desenvolvidas para analisar estes marcadores a partir de dados de NGS, tais como STRait Razor, toaSTR, e HipSTR, entre outras. Existem vários projetos colaborativos internacionais como o Projeto Genoma Humano, o projeto 1000 genomas, o *Human Genome Diversity Project* (HGDP) entre outros, que disponibilizaram os dados de sequência dos indivíduos analisados, permitindo que estes genomas possam ser estudados com diferentes abordagens voltadas para o estudo da variação genética entre populações diferentes ao redor do mundo. Esse trabalho tem como hipótese que conjuntos de marcadores STRs utilizados para identificação humana e genotipados a partir de genomas completos seriam adequados para estudos de diversidade genética e estimativas de ancestralidade populacional e individual. O objetivo geral foi avaliar os níveis de diversidade e a estrutura genética de populações humanas de diferentes regiões biogeográficas por meio de conjuntos de marcadores STR genotipados com tecnologia de sequenciamento de nova geração. Foram estudados 22 marcadores autossômicos (CSF1PO, D1S1656, D2S441, D2S1338, D3S1358, D5S818, D7S820, D8S1179, D10S1248, D12S391, D13S317, D16S539, D18S51, D19S433, D21S11, D22S1045, FGA, Penta D, Penta E, TH01, TPOX e vWA) em amostra populacional do Estado de São Paulo e em genomas sequenciados no âmbito dos projetos *1000 Genomes* e HGDP, utilizando três programas: HipSTR, STRait Razor e toaSTR. Observou-se elevada consistência e acurácia quando comparados os resultados obtidos pela utilização destas três ferramentas. Entretanto, o uso de mais de um software contribui para aumentar a acurácia na inferência de genótipos, principalmente nos marcadores (D21S11, Penta D, Penta E) em que se observou maiores taxas de erros. Com os genótipos obtidos, foi avaliada a diversidade genética entre populações de diferentes regiões biogeográficas e foi estimada a ancestralidade populacional de populações miscigenadas. O conjunto de marcadores STR aqui utilizados mostrou ser efetivos para estimar a estrutura populacional e a ancestralidade em níveis populacional e individual. Apesar da menor diversidade interpopulacional característica destes marcadores, os resultados obtidos se mostraram perfeitamente alinhados ao conhecimento pré-existente relacionado à história demográfica das populações estudadas.

Palavras-chave: Microssatélites; Repetições curtas em tandem; sequenciamento massivo em paralelo; bioinformática; genética forense.

ABSTRACT

FRONTANILLA, Tamara Soledad. Genetic diversity of global populations inferred by autosomal STRs used for human identification genotyped from complete genomes. 2022, 122 pages. [PhD Thesis]. Medicine School of Ribeirão Preto, University of São Paulo, Ribeirão Preto, São Paulo, 2023.

STR markers comprise about 3% of the human genome. One strategy to better understand the genetic make-up of a population is to estimate the distribution of allele frequencies of STR markers. The PCR technique together with capillary electrophoresis are, until today, the techniques of choice for genotyping STR markers. NGS technologies have revolutionized the biological sciences, allowing the simultaneous sequencing of many DNA or RNA samples in a short period of time. With technological advances and the increasingly frequent use of NGS techniques, there is a need to test the effectiveness of this technique in genotyping STR markers. Some tools have been developed to analyze these markers from NGS data, such as STRait Razor, toaSTR, and HipSTR, among others. There are several international collaborative projects such as the Human Genome Project, the 1000 Genomes Project, the Human Genome Diversity Project (HGDP) among others, which made available the sequence data of the analyzed individuals, allowing these genomes to be studied with different approaches aimed at studying the genetic variation among different populations around the world. The hypothesis of the present study is that sets of STR markers used for human identification and genotyped from complete genomes would be suitable for studies of genetic diversity and estimates of population and individual ancestry. The general objective was to evaluate the levels of diversity and the genetic structure of human populations from different biogeographical regions through sets of STR markers genotyped with next-generation sequencing technology. Twenty-two STR markers (CSF1PO, D1S1656, D2S441, D2S1338, D3S1358, D5S818, D7S820, D8S1179, D10S1248, D12S391, D13S317, D16S539, D18S51, D19S433, D21S11, D22S1045, FGA, Penta D, Penta E, TH01, TPOX, and vWA) were studied in a population sample from the State of São Paulo and in genomes sequenced within the scope of the 1000 Genomes and HGDP projects, using three programs: HipSTR, STRait Razor and toaSTR. High consistency and accuracy were observed when comparing the results obtained by using these three tools. However, the use of more than one software contributes to increase the accuracy in the inference of genotypes, mainly in the markers (D21S11, Penta D, Penta E) in which higher error rates were observed. With the genotypes obtained, the genetic diversity between populations from different biogeographical regions was evaluated and the population ancestry of mixed populations was estimated. The set of STR markers used here proved to be effective for estimating population structure and ancestry at population and individual levels. Despite the lower interpopulational diversity characteristic of these markers, the results obtained were perfectly aligned with the pre-existing knowledge related to the demographic history of the studied populations.

Keywords: Microsatellites; Short tandem repeats; massively parallel sequencing; bioinformatics; forensic genetics.

SUMÁRIO

LISTA DE TABELAS

Capítulo 1

Capítulo 2

Capítulo 3

LISTA DE FIGURAS

Capítulo 1

Capítulo 2

Capítulo 3

INTRODUÇÃO GERAL

1. Marcadores genéticos: identificação humana, estrutura populacional e ancestralidade

Antigamente, os principais marcadores biológicos utilizados no estudo das diferenças entre os grupos humanos eram os grupos sanguíneos e os polimorfismos proteicos. Com os avanços tecnológicos, os marcadores genéticos se destacaram devido às inúmeras vantagens sob o sistema serológico (Hosoi, 2008). Um marcador genético é um segmento de DNA com uma localização física conhecida em um cromossomo. Os marcadores mais importantes para a identificação humana e os estudos populacionais são os SNPs (do inglês, *Single-Nucleotide Polymorphisms*) e os STRs (do inglês, *Short Tandem Repeats*) (Collins, 1998).

Marcadores do tipo SNP apresentam diferenças em uma única subunidade química de DNA, ou seja, em uma única base da sequência nucleotídica. O projeto genoma humano mostrou que um genoma típico difere de um genoma de referência em 4,1 milhões a 5,0 milhões de locais e mais do 99,9% das variantes são SNPs e indels curtos. No geral, considerando os dados do Projeto 1000 Genomes, no qual foram identificados 84,7 milhões de SNPs, tais marcadores ocorrem, em média, a cada 35 bases do genoma humano (1000 Genomes Project Consortium et. al, 2015). Alguns SNP têm demonstrado ser população-específicos por apresentarem alelos com diferenças significativas em suas frequências entre os principais grupos biogeográficos. Estes marcadores foram inicialmente denominados como “alelos específicos de população” (PSAs, do inglês *Population-Specific Alleles*) ou, mais conhecidos atualmente como “marcadores informativos de ancestralidade” (AIMs, do inglês, *Ancestry Informative Markers*), com diferenças superiores a 30% nas frequências alélicas entre duas populações (Homburguer et al., 2015; Ehrlich et al., 1969; National Human Genome Research Institute, 2022). Estes marcadores são os de escolha na estimativa de ancestralidade. Outros SNPs são caracterizados por apresentar alelos bem frequentes em diferentes populações mundiais, e com pequenas diferenças nas frequências alélicas entre tais populações. Estes marcadores são úteis no processo de identificação humana, particularmente em amostras degradadas (Boonnyarit et al., 2014). Entretanto, os marcadores STRs ou microssatélites são os de escolha para esta finalidade.

Os STRs são constituídos por unidades curtas de um a seis pares de bases que se repetem de maneira adjacente, sendo observado mais de uma dezena de alelos distintos (isto é, com diferentes quantidades de repetições) em cada lócus. Tais marcadores se encontram distribuídos pelo genoma humano. Mais de meio milhão de marcadores STRs se encontram no genoma humano representando cerca de 3% do genoma e ocorrendo em média a cada 2.000 pb (Fan et al., 2007; Lander et al., 2001). Estes marcadores apresentam várias vantagens em relação aos SNPs, tais como multialelismo e alta heterozigose. Seu alto nível de polimorfismo, permite uma discriminação precisa entre indivíduos mesmo que altamente relacionados, tornando-os ótimos marcadores para identificação humana que é um dos maiores desafios na área forense (Butler, 2010). Atualmente, a maioria dos laboratórios de genética forense utilizam a técnica de PCR junto com a eletroforese capilar para estabelecer perfis genéticos de marcadores STRs que permitam a identificação humana. Esse protocolo é utilizado para comparação de amostras questionadas diretamente com amostras do suspeito, ou em casos de paternidade, desastres em massa, etc., sendo possível estabelecer um perfil a partir de baixa quantidade de DNA (0.3 ng ou até menos) (Hartl, Andrew, 2010).

Uma estratégia para compreender melhor a constituição genética de uma população é estimar a distribuição das frequências alélicas de marcadores STRs. A variação genética dentro da população é responsável pela maioria da diversidade humana (Li et al., 2009), e pode ser estudada por meio de inúmeras ferramentas de genética populacional. Uma importante consiste no modelo do equilíbrio de Hardy-Weinberg, que demonstra que as frequências genotípicas de uma população se mantêm constantes quando não há forças evolutivas atuando sobre ela, uma vez que a herança mendeliana, por si só, não gera mudança evolutiva (Beiguelman, 2008). Outro parâmetro estatístico importante é o índice de fixação (F_{ST}), que mede a diferenciação genética interpopulacional. Conhecer esses parâmetros proporciona inúmeras informações sobre a estrutura genética de populações; por exemplo, a ausência de cruzamentos aleatórios em uma população pode levar ao aumento da endogamia, enquanto que deriva genética associada a fluxo gênico reduzido pode resultar em uma alta diferenciação populacional (Hartl, 2010).

A ancestralidade de uma população pode ser estimada pela comparação de frequências alélicas com as de populações consideradas ancestrais, chamadas também populações parentais ou de referência (Pereira et al., 2019). Para o cálculo são utilizados programas computacionais e normalmente são empregados marcadores do tipo SNP, por apresentarem diferenças significativas em suas frequências entre os principais grupos biogeográficos. Na prática, a inferência de ancestralidade genética já vem sendo utilizada e se mostra importante para auxiliar nas investigações criminais desprovidas de suspeitos. A partir do DNA achado no local de crime, é possível trazer informações que se correlacionam com características fenotípicas do indivíduo procurado, contribuindo assim para diminuir o universo de busca (Koch, Andrade, 2008). Devido a que os marcadores STRs são os de escolha na área forense, e visto que a informação de ancestralidade genômica pode ser obtida diretamente a partir dos genótipos que compõem o perfil do indivíduo, são necessários mais estudos avaliando a efetividade desses marcadores para estimativa de ancestralidade e diferenciação populacional.

2. STRs e estimativa de ancestralidade

Os marcadores STRs compreendem cerca de 3% da sequência do genoma humano(8). As taxas de mutação em STRs são altas em comparação com SNPs e, portanto, representam uma grande fonte de variação genética. Estima-se que cada indivíduo abriga cerca de 100 mutações *de novo* em STRs (Fotsing et al., 2019). Os microssatélites são os marcadores recomendados para a obtenção de perfis genéticos, devido a que o número de repetições é altamente variável entre os indivíduos, o que oferece um alto poder de discriminação, tornando-os ótimos para fins de identificação (Wyner et al., 2020; Butler et al., 2007). Além disso, o pequeno tamanho dos fragmentos (100 a 200 pb) permite examiná-los a partir de amostras parcialmente degradadas ou em baixa quantidade (Eskanndarion et al., 2015).

O *Federal Bureau of Investigation* (FBI) dos Estados Unidos nomeou inicialmente 13 *loci* STRs autossômicos (CSF1PO, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, FGA, TH01, TPOX e vWA) para formar o núcleo do *Combined DNA Index System* (CODIS), o sistema que abrange o banco de dados de perfis criminais. Com o tempo foram adicionados outros sete marcadores (D1S1656, D2S441, D2S1338, D10S1248, D12S391, D19S433 e D22S1045), totalizando atualmente 20 STRs (Federal Bureau of Investigation, 2019; Hares, 2012; Hares, 2015).

Os marcadores que compõem o sistema CODIS são altamente polimórficos no mundo e caracterizados por baixa diversidade interpopulacional (F_{ST}) entre as diferentes regiões biogeográficas (Jobling, 2022). A pesar disso, é importante ressaltar que tal parâmetro é naturalmente reduzido quando se empregam marcadores que apesentam alta heterozigose (Holsinger, Weir, 2009) mas sua alta heterozigose fortalece o poder de STRs forenses para estimar a ancestralidade genômica quando comparado a outros conjuntos de STRs selecionados aleatoriamente (Holsinger, Weir, 2009; Weir et al., 2007). Como os STRs do CODIS se encontram dentro do grupo de marcadores mais analisados mundialmente é interessante avaliar a eficiência para estimativa de ancestralidade populacional.

A literatura sugere que os marcadores informativos de ancestralidade (AIMs) bialélicos geram estimativas de mistura mais acuradas e com menor desvio padrão em comparação com os STRs (Butler et al., 2007; Phillips et al.,2013). Entretanto, há a necessidade de análises que utilizem uma grande quantidade de STRs forenses (por exemplo, os 20 STRs do sistema CODIS) para avaliar sua real efetividade.

Independentemente de qual conjunto se mostrar mais informativo, a combinação de STRs e SNPs melhora a taxa de sucesso das atribuições de ancestralidade, ao mesmo tempo em que fornece um sistema confiável para detecção de mistura de perfis genéticos, o que representa uma aplicação extremamente útil em casos forenses (Phillips et al.,2013). Visto que os STRs do sistema CODIS compõem a principal ferramenta para obtenção de perfil criminal de DNA, se fazem necessários mais estudos acerca da diversidade genética global destes marcadores, bem como da efetividade destes marcadores na diferenciação de populações.

3. Genotipagem de marcadores STRs

A técnica de PCR junto com a eletroforese capilar (EC) são, até hoje, as técnicas de escolha para a genotipagem de marcadores STR, sendo utilizadas na maioria dos laboratórios forenses do mundo (Eskandarion et al., 2015). A PCR permite a produção (amplificação) rápida de bilhões de fragmentos de um segmento específico de DNA e a eletroforese capilar utiliza uma corrente elétrica para separar essas moléculas com base em seu tamanho (Garibyan et al., 2013). Vários kits comerciais foram desenvolvidos para amplificação dos STRs utilizados para identificação humana incluindo os marcadores estabelecidos pelo CODIS (Morling, 2009; Kraemer et al., 2017). Thermo Fisher Scientific, Promega e Qiagen possuem diversos kits desenvolvidos especificamente para casos de identificação humana. Os kits GlobalFiler (Thermo Fisher Scientific) (Wang et al., 2015; Flores et al., 2014; Applied Biosystems, 2021) e PowerPlex Fusion System (Promega) (Ensenberger et al., 2016; Ludeman et al., 2018) são conjuntos mais recentes que atendem aos requerimentos europeus e norte-americanos (CODIS), sendo amplamente utilizados na rotina diária dos laboratórios forenses (Kraemer et al., 2017; Cho et al.,2021).

Em 2000 e empresa Lynx Therapeutics lançou a primeira das tecnologias NGS (NGS, do inglês, *Next Generation Sequencing*) – termo utilizado para qualquer metodologia que não era baseada no método de sequenciamento de Sanger (Barba et al., 2014). Várias empresas foram publicando suas tecnologias em tempos similares. Apesar das diferencias substanciais entre elas, todas têm como principal característica o processamento paralelo massivo de fragmentos de DNA. Enquanto um sequenciador que utiliza eletroforese capilar processa, no máximo, 96 fragmentos por vez (Slatko et al.,2018; Thermo Fisher Scientific,), os sequenciadores de nova geração, como Ion Torrent (Thermo Fisher Scientific), e Illumina Inc., podem ler até bilhões de fragmentos simultaneamente. Tais plataformas vêm revolucionando o processo de genotipagem de STRs (Smetana et al., 2022; Lahens et al., 2017).

De fato, as tecnologias de NGS revolucionaram as ciências biológicas. Permitem o sequenciamento simultâneo de muitas amostras de DNA ou RNA em um curto período de tempo. Além disso, possibilita analisar vários alvos simultaneamente e reconhecer locais de variação ou mutações no genoma (Koboldt et al., 2013). Na maioria das metodologias de NGS, o DNA é quebrado em pequenos fragmentos, que são sequenciados e depois realinhados através de ferramentas bioinformáticas utilizando a sequencia de um genoma de referência. As técnicas de NGS mais difundidas em laboratórios forenses permitem analisar fragmentos de até 300 pares de bases (Zhang et al., 2014). Embora existam tecnologias voltadas para sequenciar longos fragmentos (em torno de 10 mil a 1000 mil pb), a aplicação de tais tecnologias à genética forense é limitada devido à natureza degradada de muitas amostras forenses.

Inicialmente o problema da técnica de NGS era o custo elevado, limitando assim o acesso a essa técnica. Segundo os dados do *National Human Genome Research Institute*, em 2001 o custo para o sequenciamento de um genoma humano era de cerca de 100 milhões de dólares. Porém, os avanços no campo da genômica no último quarto de século levaram a reduções substanciais no custo, permitindo hoje em dia sequenciar um genoma por menos de 1000 dólares (Wetterstrand, 2014).

Com os avanços tecnológicos e a utilização cada vez mais frequente das técnicas de NGS surge a necessidade de testar a efetividade dessa técnica na genotipagem de marcadores STR (Fungtammasann et al., 2015; Bornman et al., 2012). Porém, esse processo tem sido desafiador devido às altas taxas de erro no alinhamento e presença de artefatos (*stutters*). Entretanto, estudos recentes demostraram que com sequenciamento de alta cobertura e a utilização de programas específicos, os marcadores STR podem ser genotipados de maneira exitosa (Valle-Silva et al., 2022; Willems et al., 2017; Gymrek et al., 2012; Ganschow et al., 2018; Warshauer et al., 2013).

Outra grande vantagem da técnica de NGS na área forense é que permitiria a diferenciação de isoalelos (alelos com mesmo comprimento, porém, com diferentes sequências), o que não é possível com a técnica de eletroforese capilar e aumentaria ainda mais o poder de diferenciação entre indivíduos. A desvantagem dessa técnica é que ainda são necessárias ferramentas bioinformáticas muitas vezes complexas ou um alto nível de conhecimento em programação para a obtenção e análise dos genótipos, o que pode limitar a utilização dessa tecnologia.

Mesmo que os resultados das pesquisas atuais sejam favoráveis, o desenvolvimento de novas ferramentas bioinformáticas para a obtenção e análise de genótipos de STRs a partir de sequenciamento de nova geração é imprescindível.

4. Ferramentas bioinformáticas para genotipagem de marcadores STR a partir de dados de sequenciamento de nova geração

Algumas ferramentas foram desenvolvidas para analisar marcadores STR a partir de dados de NGS, tais como STRait Razor (Warshauer et al., 2013), toaSTR (Ganschow et al., 2018), e HipSTR (Willems et al., 2017; Willems et al., 2014), entre outras. Cada software utiliza diferentes algoritmos e regiões franqueadoras para capturar os STR. Abordaremos estas três ferramentas por serem gratuitas e porque se mostraram efetivas em estudos prévios.

STRait Razor (Warshauer et al., 2013) consiste em uma ferramenta simples que funciona como uma extensão do programa Microsoft Excel e permite o processamento de várias amostras simultaneamente. Para isso, são necessários os dados FASTq das duas leituras; R1 e R2, e como resultado da análise gera uma planilha com todas as sequências encontradas para cada marcador, incluindo as respectivas coberturas individuais. Posteriormente, é necessária uma análise manual para avaliar as coberturas e, assim, diferenciar os alelos de possíveis artefatos (*stutters*).

Ao contrário do STRait Razor, o toaSTR (Ganschow et al., 2018) permite o processamento de apenas uma amostra por vez. Em seus primórdios era uma ferramenta on-line, porém hoje em dia funciona com uma máquina virtual chamada Docker. São necessários os arquivos BAM das duas leituras; R1 e R2. Como output o programa oferece arquivos independentes de cada leitura, incluindo os alelos encontrados e as coberturas. Estes dados devem ser analisados de forma conjunta para descartar artefatos ou erros de genotipagem.

Por fim, o software HipSTR (Willems et al., 2017) foi especificamente desenvolvido para o sequenciamento Illumina, buscando diminuir os erros de genotipagem observados em ferramentas anteriores. Ele funciona por meio de linhas de comando e precisa dos arquivos BAM ou FASTq das duas leituras; R1 e R2. É um programa um pouco mais complexo, demandando maior conhecimento de bioinformática. O HipSTR gera um arquivo VCF que contém todas as informações relevantes para determinação dos alelos STRs em uma única linha. Ao contrário do STRait Razor e do toaSTR, o HipSTR não fornece os alelos diretamente. Em vez disso, informa as inserções ou deleções em termos do número de bases para cada alelo em relação ao alelo de referência para cada marcador. Uma grande vantagem do HipSTR é que permite mudar as regiões flanqueadoras; essa característica pode ser de grande utilidade para a genotipagem de marcadores que possuem comprimento muito grande, uma vez que possibilita alternativas em cenários onde os *reads* não incluem as duas regiões flanqueadoras, ou ao menos a região de repetição por completo.

Todos os programas se mostraram efetivos (Valle-Silva et al., 2022), porém cada um deles tem diferentes vantagens e desvantagens que devem ser a avaliadas antes de escolher qual programa utilizar. Fatores como tipo de amostra, formato de arquivo de entrada disponível (i.e., FASTq, BAM ou CRAM), quantidade de amostras a serem analisadas, dentre outros, são determinantes. Além disso, as ferramentas bioinformáticas em geral se encontram em constante atualização e, portanto, são necessários mais estudos avaliando estas e possíveis novas ferramentas que venham a ser desenvolvida para a finalidade de genotipagem de STRs.

5. Projetos colaborativos internacionais

Existem vários projetos colaborativos internacionais como o Projeto Genoma Humano (Lander et al., 2001), o projeto 1000 genomas (Smetana et al., 2022), o projeto HGDP (Cavalli-Sforza, 2005) entre outros, que disponibilizaram os dados de sequência dos indivíduos analisados, permitindo que estes genomas possam ser estudados com diferentes abordagens voltadas para o estudo da variação genética entre diferentes populações ao redor do mundo.

O projeto *Human Genome Diversity Project* ou HGDP (Cavalli-Sforza, 2005) foi iniciado pelo Instituto Morrison da Universidade de Stanford em 1990. O projeto se baseia em amostras de populações autóctones coletadas ao redor do mundo, muitas das quais são consideradas como populações em risco de desaparecimento. Por conta disso, nas suas fases iniciais enfrentou vários problemas éticos; porém, após 4 anos de discussão, o *US National Research Council* (NRC) da *National Academy of Sciences* (NAS) recomendou que o projeto continuasse com vários e cuidadosos protocolos éticos devido aos grandes benefícios da pesquisa. O projeto HGDP estudou 54 populações diferentes distribuídas ao redor do mundo, abordando sete grupos populacionais: África, Américas (ameríndios), Centro-Sul Asiático, Leste Asiático, Europa, Oriente Médio e Oceania. As sequências dos genomas avaliados estão disponíveis para estudo no portal do International Genome Sample Resource (IGSR): https://www.internationalgenome.org/data-portal/data-collection.

O projeto *1000 Genomes* (Smetana et al., 2022) foi uma colaboração mundial que produziu um extenso catálogo de variação genética humana. Inicialmente foram sequenciados genomas inteiros de 2.504 indivíduos divididos em cinco grandes grupos: África, leste da Ásia, Europa, sul da Ásia e Américas (americanos miscigenados). A abordagem inicial envolveu sequenciamento com baixa cobertura (~7.4x), o que limitou sua utilização (1000 Genomes Project Consortium et al., 2015). Recentemente o *New York Genome Center* sequenciou novamente as 2.504 amostras com alta cobertura (30x). Além disso, alinharam os dados de sequência em relação ao genoma de referência GRCh38. Os dados gerados estão publicamente disponíveis no portal do International Genome Sample Resource (IGSR): https://www.internationalgenome.org/data-portal/data-collection/30x-grch38.

Estes projetos internacionais não tiveram como foco principal estudar marcadores do tipo STR. Adicionalmente, os avanços na área de genômica e a diminuição do custo, tornaram possível a utilização da técnica de sequenciamento de genoma completo (WGS) de forma mais acessível. Atualmente, muitos pesquisadores estão realizando WGS para conhecer, por exemplo, probabilidades e riscos de desenvolver doenças poligênicas e multifatoriais, o que seria mais trabalhoso e caro usando metodologias tradicionais. Levando isso para a área forense, onde os marcadores STRs são os de escolha para identificação humana, esses dados já sequenciados poderiam ser utilizados como um banco de dados de frequências populacionais (West et al., 2020; Carratto et al., 2022), tendo aplicabilidade em exames de vínculos genéticos e testes de paternidade (Borstinng et al., 2015; Alvarez-Cubero et al., 2017). No futuro, com o melhoramento das ferramentas bioinformáticas, seria possível a identificação imediata de isoalelos, o que é impossível com técnicas tradicionais como a PCR associada a eletroforese capilar (Alvarez-Cubero et al., 2017; Ballard et al., 2020). Isso seria uma grande diferencial e aumentaria ainda mais o poder de discriminação dos marcadores STR usados hoje em dia. Portanto, mais estudos são necessários para otimizar recursos e conseguir avaliar de uma forma simples a grande quantidade de dados gerados pela tecnologia de NGS.

HIPÓTESE

Conjuntos de marcadores STRs utilizados para identificação humana podem ser genotipados com acurácia a partir de genomas completos gerados por metodologias voltadas para sequenciamento de pequenos fragmentos (*short reads*) e seriam adequados para estudos de diversidade genética e estimativas de ancestralidade populacional e individual.

OBJETIVOS

Objetivo geral

Avaliar os níveis de diversidade e a estrutura genética de populações humanas de diferentes regiões biogeográficas por meio de conjuntos de marcadores STR genotipados com tecnologia de sequenciamento de nova geração.

Objetivos específicos

<u>Capítulo 1</u>

- Avaliar a concordância das ferramentas HipSTR, STRait Razor e toaSTR para genotipar marcadores STR a partir de dados de NGS de uma amostra populacional brasileira altamente miscigenada.

<u>Capítulo 2</u>

- Analisar o conjunto de STRs autossômicos genotipados com a ferramenta HipSTR utilizando os dados de genomas sequenciados no contexto do Projeto Diversidade do Genoma Humano (HGDP).
- Validar a utilização da ferramenta HipSTR pela comparação dos genótipos aqui obtidos com genótipos determinados nas mesmas amostras por PCR e eletroforese capilar e disponibilizados em banco de dados de Rosenberg e colaboradores (2016).

- Publicar um banco de dados global de genótipos e frequências alélicas de STRs autossômicos baseado em todas a populações do projeto HGDP.

Capítulo 3

- Realizar uma análise populacional abrangente baseada na genotipagem de STRs autossômicos utilizados na área forense utilizando-se a ferramenta HipSTR a partir de dados de genomas sequenciados no contexto do projeto 1000 Genomas.
- Publicar um banco de dados global de frequências alélicas de STRs autossômicos baseado em todas a populações do projeto 1000 Genomas.

REFERÊNCIAS BIBLIOGRÁFICAS

Alvarez-Cubero M.J, Saiz M, Martínez-García B, Sayalero S.M, Entrala C, Lorente J.A, Martinez-Gonzalez L.J (2017). Next generation sequencing: an application in forensic sciences? Ann Hum Biol. 44(7):581-592. Doi: 10.1080/03014460.2017.

Applied biosystems (2021). Integrated human identification (HID) solutions. p.7. Avaliable at: https://assets.thermofisher.com/TFS-Assets/GSD/brochures/hid-integrated-solutions-brochure.pdf

Ballard D, Winkler-Galicki J, Wesoły J (2020). Massive parallel sequencing in forensics: advantages, issues, technicalities, and prospects. Int J Legal Med 134(4):1291-1303. Doi: 10.1007/s00414-020-02294-0.

Barba M, Czosnek H, Hadidi A (2014). Historical perspective, development and applications of next-generation sequencing in plant virology. Viruses. 6;6(1):106-36. Doi: 10.3390/v6010106.

Beiguelman B. Genética de Populações humanas: Editora SBG Sociedade Brasileira de Genética, 2008. 239p.

Boonyarit H, Mahasirimongkol S, Chavalvechakul N, Aoki M, Amitani H, Hosono N, Kamatani N, Kubo M, Lertrit P (2014). Development of a SNP set for human identification: A set with high powers of discrimination which yields high genetic information from naturally degraded DNA samples in the Thai population. Forensic Sci Int Genet. 11:166-73. Doi: 10.1016/j.fsigen.2014.03.010.

Bornman, D.M.; Hester, M.E.; Schuetter, J.M.; Kasoji, M.D.; Minard-Smith, A.; Barden, C.A.; Nelson, S.C.; Godbold, G.D.; Baker, C.H.; Yang, B.; et al (2012). Short-read, high-throughput sequencing technology for STR genotyping. Biotech. Rapid Dispatches 1–6. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4301848/

Børsting, C.; Morling, N (2015). Next generation sequencing and its applications in forensic genetics. Forensic Sci. Int. Genet. 18, 78–89. Doi: 10.1016/j.fsigen.2015.02.002

Butler J. Fundamentals of Forensic DNA Typing: Academic Press; 2010, 500p.

Butler J.M, Coble M.D, Vallone P.M (2007). STRs vs. SNPs: thoughts on the future of forensic DNA testing. Forensic Sci Med Pathol. 3(3):200-5. Doi: 10.1007/s12024-007-0018-1.

Carratto, T.M.T.; Moraes, V.M.S.; Recalde, T.S.F.; Oliveira, M.L.G.; Teixeira Mendes-Junior, C (2022). Applications of massively parallel sequencing in forensic genetics. Genet. Mol. Biol. 45, e20220077. Doi: 10.1590/1678-4685-GMB-2022-0077

Cavalli-Sforza LL (2005). The Human Genome Diversity Project: past, present and future. Nat Rev Genet. 6(4):333-40. Doi: 10.1038/nrg1596.

Cho W.C, Jung J.K, Cho Y, Jung J.Y, Lee M.H, Park J.H, Lee D.S, Lee J (2021). Validation and assessment of the Investigator® 24plex QS kit for forensic casework application: Comparison with the PowerPlex® fusion system and GlobalFiler™ PCR amplification kits. Leg Med (Tokyo). 52:101902. Doi: 10.1016/j.legalmed.2021.101902.

Collins FS, Brooks LD, Chakravarti A (1998). A DNA polymorphism discovery resource for research on human genetic variation. Genome Res. 8(12):1229-31. Doi: 10.1101/gr.8.12.1229. Erratum in: Genome Res 1999 Feb;9(2):210

Ehrlich PR, Raven PH (1969). Differentiation of populations. Sciencem19;165(3899):1228-32. Doi: 10.1126/science.165.3899.1228.

Ensenberger M.G, Lenz K.A, Matthies L.K, Hadinoto G.M, Schienman J.E, Przech A.J, et al. (2016). Developmental validation of the PowerPlex(®) Fusion 6C System. Forensic Sci Int Genet. 21:134-44. Doi: 10.1016/j.fsigen.2015.12.011.

Eskandarion M, Najafi M, Akbari Eidgahi M, Alipour Tabrizi A, Golmohamadi T. Optimization of short tandem repeats (STR) typing method and allele frequency of 8 STR markers in referring to forensic medicine of Semnan Province. J Med Life. 2015;8(Spec Iss 4):180-185. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5319289/

Fan H.; Chu J.Y (2007). A brief review of short tandem repeat mutation. Genom. Proteom. Bioinform. 5(1):7–14. Doi: 10.1016/S1672-0229(07)60009-6

Federal Bureau of Investigation. Frequently Asked Questions on CODIS and NDIS. Available at: https://www.fbi.gov/how-we-can-help-you/dna-fingerprint-act-of-2005-expungement-policy/codis-and-ndis-fact-sheet

Flores S, Sun J, King J, Budowle B (2014). Internal validation of the GlobalFiler™ Express PCR Amplification Kit for the direct amplification of reference DNA samples on a high-throughput automated workflow. Forensic Sci Int Genet. 10:33-39. Doi: 10.1016/j.fsigen.2014.01.005.

Fotsing SF, Margoliash J, Wang C, Saini S, Yanicky R, Shleizer-Burko S, Goren A, Gymrek M (2019). The impact of short tandem repeat variation on gene expression. Nat Genet. 51(11):1652-1659. doi: 10.1038/s41588-019-0521-9.

Fungtammasan, A.; Ananda, G.; Hile, S.E.; Su, M.S.; Sun, C.; Harris, R.; Medvedev, P.; Eckert, K.; Makova, K.D (2015). Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. Genome Res. 25, 736–749. Doi: 10.1101/gr.185892.114.

Ganschow, S.; Silvery, J.; Kalinowski, J.; Tiemann, C (2018). toaSTR: A web application for forensic STR genotyping by massively parallel sequencing. Forensic Sci. Int. Genet. 37, 21–28. Doi: 10.1016/j.fsigen.2018.07.006.

Garibyan L, Avashia N (2013). Polymerase chain reaction. J Invest Dermatol. 133(3):1-4. Doi: 10.1038/jid.2013.1.

Gymrek, M.; Golan, D.; Rosset, S.; Erlich, Y (2012). lobSTR: A short tandem repeat profiler for personal genomes. Genome Res. 22, 1154–1162. Doi: 10.1101/gr.135780.111.

Hares DR (2012). Expanding the CODIS core loci in the United States. Forensic Sci Int Genet. 6(1):e52-4. doi: 10.1016/j.fsigen.2011.04.012. Erratum in: Forensic Sci Int Genet. 2012 Sep;6(5):e135.

Hares DR. (2015) Selection and implementation of expanded CODIS core loci in the United States. Forensic Sci Int Genet. 17:33-4. Doi: 10.1016/j.fsigen.2015.03.006

Hartl D, Andrew C. Principios de genética de populações. 4th edition ed2010. 660p.

Holsinger KE, Weir BS (2009). Genetics in geographically structured populations: defining, estimating and interpreting F(ST). Nat Rev Genet. 10(9):639-50. Doi: 10.1038/nrg2611.

Homburger JR, Moreno-Estrada A, Gignoux CR, Nelson D, Sanchez E, Ortiz-Tello P, Pons-Estel BA, Acevedo-Vasquez E, Miranda P, Langefeld CD, Gravel S, Alarcón-Riquelme ME, Bustamante CD (2015). Genomic Insights into the Ancestry and Demographic History of South America. PLoS Genet. 4;11(12):e1005602. Doi: 10.1371/journal.pgen.1005602.

Hosoi E (2008). Biological and clinical aspects of ABO blood group system. J Med Invest. 55(3-4):174-82. Doi: 10.2152/jmi.55.174.

Illumina (2022). Introduction to NGS. Available from: https://www.illumina.com/science/technology/next-generation-sequencing.html

Jobling, M.A (2022). Forensic genetics through the lens of Lewontin: Population structure, ancestry and race. Philos. Trans. R. Soc. Lond. B Biol. Sci. 2022, 377, 20200422. Doi: 10.1098/rstb.2020.0422

Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER (2013). The next-generation sequencing revolution and its impact on genomics. Cell. 26;155(1):27-38. Doi: 10.1016/j.cell.2013.09.006.

Koch A, Andrade F (2008). A utilização de técnicas de biologia molecular na genética forense: uma revisão. RBAC 40(1):17-23. Available from: https://moodle.ufsc.br/file.php/19765/topico_vii/genetica_forense-1.pdf

Kraemer M, Prochnow A, Bussmann M, Scherer M, Peist R, Steffen C (2017). Developmental validation of QIAGEN Investigator® 24plex QS Kit and Investigator® 24plex GO! Kit: Two 6-dye multiplex assays for the extended CODIS core loci. Forensic Sci Int Genet. 29:9-20. Doi: 10.1016/j.fsigen.2017.03.012.

Lahens NF, Ricciotti E, Smirnova O, Toorens E, Kim EJ, Baruzzo G, Hayer KE, Ganguly T, Schug J, Grant GR (2017). A comparison of Illumina and Ion Torrent sequencing platforms in the context of differential gene expression. BMC Genomics.10;18(1):602. Doi: 10.1186/s12864-017-4011-0.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, et al (2001). Initial sequencing and analysis of the human genome. Nature. 15;409(6822):860-921. Doi: 10.1038/35057062. Erratum in: Nature 2001 Aug 2;412(6846):565. Erratum in: Nature 2001 Jun 7;411(6838):720.

Li H, Tang H, Zhang Q, Jiao Z, Bai J, Chang S (2009). A multiplex PCR for 4 X chromosome STR markers and population data from Beijing Han ethnic group. Leg Med (Tokyo). 11(5):248-50. Doi: 10.1016/j.legalmed.2009.03.013.

Morling N (2009). PCR in forensic genetics. Biochem Soc Trans.37(Pt 2):438-40. Doi: 10.1042/BST0370438.

National Human Genome Research Institute (2022). Ancestry Informative Markers. Available from: https://www.genome.gov/genetics-glossary/Ancestry-informative-Markers#:~:text=Definition&text=Ancestry%2Dinformative%20markers%20are%20sets,geographical%20regions%20of%20the%20world.

Parson W, Ballard D, Budowle B, Butler JM, Gettings KB, Gill P, et al (2016). Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements. Forensic Sci Int Genet.22:54-63. Doi: 10.1016/j.fsigen.2016.01.009

Pereira FDSCF, Guimarães RM, Lucidi AR, Brum DG, Paiva CLA, Alvarenga RMP (2019). A systematic literature review on the European, African and Amerindian genetic ancestry components on Brazilian health outcomes. Sci Rep. 20;9(1):8874. Doi: 10.1038/s41598-019-45081-7. Erratum in: Sci Rep. 2020 May 1;10(1):7677.

Phillips C, Fernandez-Formoso L, Gelabert-Besada M, Garcia-Magariños M, Santos C, Fondevila M, Carracedo A, Lareu MV (2013). Development of a novel forensic STR multiplex for ancestry analysis and extended identity testing. Electrophoresis. 34(8):1151-62. Doi: 10.1002/elps.201200621.

Promega (2022). PowerPlex® Fusion System. Available from: https://worldwide.promega.com/products/forensic-dna-analysis-ce/str-amplification/powerplex-fusion-system/?catNum=DC2402

Slatko BE, Gardner AF, Ausubel FM (2018). Overview of Next-Generation Sequencing Technologies. Curr Protoc Mol Biol. 122(1):e59. Doi: 10.1002/cpmb.59.

Smetana J, Brož P (2022). National Genome Initiatives in Europe and the United Kingdom in the Era of Whole-Genome Sequencing: A Comprehensive Review. Genes (Basel) 21;13(3):556. DDoi: 10.3390/genes13030556.

Thermo Fisher Scientific (2022). Ion Torrent next-generation sequencing. Available from: https://www.thermofisher.com/py/en/home/brands/ion-torrent.html.

Valle-Silva, G.; Frontanilla, T.S.; Ayala, J.; Donadi, E.A.; Simões, A.L.; Castelli, E.C.; Mendes-Junior, C.T (2022). Analysis and comparison of the STR genotypes called with HipSTR, STRait Razor and toaSTR by using next generation sequencing data in a Brazilian population sample. Forensic Sci. Int. Genet. 58, 102676. Doi: 10.1016/j.fsigen.2022.102676

Wang DY, Gopinath S, Lagacé RE, Norona W, Hennessy LK, Short ML, Mulero JJ (2015). Developmental validation of the GlobalFiler(®) Express PCR Amplification Kit: A 6-dye multiplex assay for the direct amplification of reference samples. Forensic Sci Int Genet. 19:148-155. Doi: 10.1016/j.fsigen.2015.07.013.

Warshauer, D.H.; Lin, D.; Hari, K.; Jain, R.; Davis, C.; Larue, B.; King, J.L.; Budowle, B (2013). STRait Razor: A length-based forensic STR allele-calling tool for use with second generation sequencing data. Forensic Sci. Int. Genet. 7, 409–417. Doi: 10.1016/j.fsigen.2013.04.005.

Weir BS (2007). The rarity of DNA profiles. Ann Appl Stat. 1(2):358-370. Doi: 10.1214/07-AOAS128.

West, F.L.; Algee-Hewitt, B.F.B (2020). Cadaveric blood cards: Assessing DNA quality and quantity and the utility of STRs for the individual estimation of trihybrid ancestry and admixture proportions. Forensic Sci. Int. Synerg. 2, 114–122. Doi: 10.1016/j.fsisyn.2020.03.002.

Wetterstrand, K.A. (2014) DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Available from: www.genome.gov/sequencingcosts

Willems, T.; Gymrek, M.; Highnam, G.; Mittelman, D.; Erlich, Y.; Consortium, G.P (2014). The landscape of human STR variation. Genome Res. 24, 1894–1904. Doi: 10.1101/gr.177774.114.

Willems, T.; Zielinski, D.; Yuan, J.; Gordon, A.; Gymrek, M.; Erlich, Y (2017). Genome-wide profiling of heritable and de novo STR variations. Nat. Methods 14, 590–592. Doi: 10.1038/nmeth.4267

Wyner N, Barash M, McNevin D (2020). Forensic Autosomal Short Tandem Repeats and Their Potential Association With Phenotype. Front Genet. 6(11):884. Doi: 10.3389/fgene.2020.00884.

Zhang W, Cui H, Wong LJ (2014). Application of next generation sequencing to molecular diagnosis of inherited diseases. Top Curr Chem. 336:19-45. Doi: 10.1007/128_2012_325.

1000 Genomes Project Consortium, Auton A, Brooks L.D, Durbin R.M, Garrison E.P, Kang H.M, Korbel J.O, Marchini J.L, McCarthy S, McVean G.A, Abecasis G.R 2015. A global reference for human genetic variation. Nature 1;526(7571):68-74. Doi: 10.1038/nature15393.

CAPÍTULO 1

CAPÍTULO 1

Analysis and comparison of the STR genotypes called with HipSTR, STRait Razor and toaSTR by using next generation sequencing data in a Brazilian population sample

Guilherme Valle-Silva[a,1], Tamara Soledad Frontanilla[b,1,] Jesús Ayala[c], Eduardo Antonio Donadi[d], Aguinaldo Luiz Simo˜es[b], Erick C. Castelli[e], Celso Teixeira Mendes-Junior[a]

[a] *Departamento de Química, Laboratório de Pesquisas Forenses e Genômicas, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de Sa˜o Paulo, 14040-901 Ribeirão Preto, SP, Brazil.*

[b] *Departamento de Genética, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, 14049-900 Ribeirão Preto, SP, Brazil.*

[c] *Softec S.R.L. Asunción, Paraguay.*

[d] *Divisão de Imunologia Clínica, Departamento de Clínica Médica, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, 14048-900 Ribeira˜o Preto, SP, Brazil.*

[e] *São Paulo State University (UNESP), Department of Pathology, School of Medicine, Botucatu, SP, Brazil.*

[1] These authors contributed equally to this work.

Doi: https://doi.org/10.1016/j.fsigen.2022.102676

Supplementary Materials:
https://www.sciencedirect.com/science/article/pii/S1872497322000175#sec0065

Received 25 April 2021;
Received in revised form 18 January 2022;
Accepted 2 February 2022;
Available online 4 February 2022;

ABSTRACT

Short tandem repeats (STRs) are particularly difficult to genotype with rapid evolving next-generation sequencing (NGS) technology. Long amplicons containing repetitive sequences result in alignment and geno- typing errors. Stutters arising from polymerase slippage often result in reads with additional or missing repeat copies. Many tools are available for analysis of STR markers from NGS data. This study has evaluated the concordance of the HipSTR, STRait Razor, and toaSTR tools for STR genotype calling; NGS data obtained from a highly genetically diverse Brazilian population sample have been used. We found that toaSTR can retrieve a larger number of genotypes (93.8%), whereas HipSTR (84.9%) and STRait Razor present much lower genotype calling (75.3%). Accuracy levels for genotype calling are very similar (identical genotypes ~95% and correct alleles ~ 97.5%) across the three methods. All the markers presenting the same genotype through the methods are in Hardy–Weinberg equilibrium. We found that combined match probability and combined exclusion power are 2.90×10^{-28} and 0.99999999982, respectively. Although toaSTR has varying locus-specific differences and better overall performance of toaSTR, the three programs are reliable genotyping tools. Notwithstanding, additional effort is necessary to improve the genotype calling accuracy of next-generation sequencing datasets.

Keywords: Short tandem repeats; CODIS; Brazil; Massively parallel sequencing Bioinformatics; Forensic genetics

1. INTRODUCTION

Next-generation sequencing technology (NGS), or massively parallel sequencing, has revolutionized biological sciences. It allows simultaneous sequencing of many DNA and RNA samples in a short period. Besides, NGS costs have decreased significantly over the past decade (Qin, 2019). NGS allows analyzing several targets to be analyzed simultaneously and recognizes variation sites or mutations in the genome (Koboldt, et al., 2013).

Short tandem repeats (STRs) are highly repetitive genomic sequences. The human genome presents thousands of STRs. Their codominance, multiallelism, and high heterozygosity are advantageous for human identification (Fan et al., 2007). Their high level of polymorphism enables closely related individuals to be precisely discriminated, so they are the markers of choice for human identification purposes (Chen et al., 2014). The multiplex PCR technique, along with capillary electrophoresis, is currently the elective method in most forensic genetics' laboratories. However, this technique has some limitations; for example, alleles with different sequences cannot be distinguished when they present the same length (Durney et al., 2015).

With NGS, obtaining STR profiles from genome sequences has become practical. STRs are particularly challenging to genotype in NGS data. The earlier phases of NGS were characterized by short reads, so they failed to cover longer markers and their flanking regions (Fungtammasan et al., 2015; Bornman et al., 2012). Advancement in technologies provided reads with higher quality and sizes, being more suitable for STRs analyses. Notwithstanding, the STR repetitive sequences accounting for their high mutation rates also cause frequent alignment errors that can complicate and compromise genotype calling. Moreover, stutter errors often result in reads that contain additional or missing repeat copies (Willems et al., 2017). There are many programs for analysis of STR markers from NGS data, including STRait Razor (Warshauer et al., 2013), toaSTR (Ganschow et al., 2018), and HipSTR (Willems et al., 2017).

While most of the available forensic solutions, such as the Precision ID GlobalFilerTM NGS STR Panel v2 (ThermoFisher) and Forenseq DNA Signature Prep kit (Illumina), relies upon amplicon sequencing, many alternative NGS targeted enrichment methods, such as HaloPlex (Agilent Technologies), SureSelect (Agilent Technologies), and SeqCap (Roche) uses hybridization capture. In the HaloPlex technology, restriction enzymes are used to fragment genomic DNA, which are circularized by hybridization to probes whose ends are complementary to the target fragments (Berglund et al., 2013). Although amplicon-based methods are simpler and utilize smaller DNA inputs, hybridization capture assays are associated with larger target panels and usually provides higher uniformity (Samorodnitsky et al., 2015).

There are many software to analyze STR markers from NGS data, like STRait Razor (Warshauer et al., 2013), toaSTR (Ganschow et al., 2018), HipSTR (Willems et al., 2017), lobSTR (Gymrek et al., 2012), Gene-MarkerHTS (Hendricks et al.,2017), STRScan (Tang et al., 2017) and others.

The short tandem repeat allele identification tool (STRait Razor) characterizes STRs in NGS data. STRait Razor allows several samples to be analyzed simultaneously, generating an Excel spreadsheet with all the sequences found for each marker, with their respective coverages. After that, a manual analysis step must be performed to select the final genotypes of each sample (Warshauer et al., 2013).

toaSTR is a web tool for STR allele calling in NGS data. It does not depend on which sequencing platform and forensic kit are used. It allows marker panels to be customized, stutter models to be selected, and data to be visualized. Nevertheless, it only analyzes one sample per time. Like STRait Razor, it records all haplotypes detected in the sample and their respective individual coverages, and it also requires manual analysis to determine the genotypes for each marker (Warshauer et al., 2013).

The haplotype inference and phasing for STRs (HipSTR) was developed for Illumina sequencing to mitigate the current errors observed in previous STR tools. STRait Razor and toaSTR were made for independent finding of the true length of repeats within each locus, but HipSTR considers the whole repeat structure of the allele (Willems et al., 2017). The output of HipSTR is a VCF file that consolidates all the relevant information of the determination of the STR allele in a single line. In contrast to STRait Razor and toaSTR, HipSTR does not provide the alleles directly. It rather informs the insertions or deletions in terms of the number of bases for each allele with respect to the reference allele for each marker (Willems et al., 2014; Willems et al., 2017).

These three tools were chosen for this study since they are freely available, accessible, and showed very good results in previous studies (Willems et al, 2017;

Warshauer et al., 2013; Ganschow et al., 2018). STRait Razor and toaSTR are also among the most explored ones and, despite being older, they have been widely validated and are being constantly updated with the inclusion of new functionalities. HipSTR is a more complex tool that requires bioinformatics knowledge and was included in this essay because it is the new and improved version of lobSTR, being developed specifically for Illumina assays. Moreover, HipSTR employs a different capture strategy that relies upon the whole STR genome location rather than predefined flanking regions. Considering the possible application of rapidly evolving technologies for human identification purposes, this study has evaluated the concordance of the HipSTR, STRait Razor and toaSTR tools for STR genotype calling by using NGS data obtained from a highly genetically diverse Brazilian population sample.

2. METHODOLOGY

2.1 Sample and sequencing

The study was approved by the institutional Ethics Committee (Comitê de Ética em Pesquisa, FFCLRP-USP) according to protocol CAAE #25696413.7.0000.5407. The sample consisted of 547 Brazilian individuals, 279 women and 268 men, with ages ranging from 18 to 80 years, from the city of Ribeirão Preto and surrounding regions, Southeastern Brazil.

2.2 Laboratory analysis

DNA extraction was performed by using the salting-out protocol (Miller et al., 1988). Integrity, purity and concentration were assessed in the genomic DNA samples by Agarose Gel Electrophoresis, NanoDrop spectrophotometry (Thermo Fisher Scientific, Waltham, Massachusetts), and QubitTM dsDNA BR fluorimetric assay (Thermo Fisher Scientific), respectively. To achieve an ideal concentration for DNA sequencing-library preparation, all the samples were normalized to a concentration of 5 ng/µL.

A personalized HaloPlex Target Enrichment System kit (Agilent Technologies, Santa Clara, California) assay was applied to prepare the DNA library. The panel of probes was designed with the SureDesign tool (Agilent Technologies); using the hg19/GRCh37 human genome release was used a reference. Twenty-two STR markers (CSF1PO, D1S1656, D2S441, D2S1338, D3S1358, D5S818, D7S820, D8S1179, D10S1248, D12S391, D13S317, D16S539, D18S51, D19S433, D21S11, D22S1045, FGA, TH01, TPOx, vWA, Penta D, and Penta E), as well as other regions of interest were analyzed for different studies. The designed probes (Supplementary Table 1) were checked with the UCSC Genome Browser to confirm targeting regions.

According to the manufacturer's instructions, 5 µL of each sample was digested by eight different pairs of enzymes to create libraries of DNA fragments. HaloPlex biotinylated probes were used to capture fragments, and indices were incorporated for sample identification. The captured fragments were amplified by PCR; Herculase II Fusion polymerase was employed in the SureCycle 8800 thermocycler (Agilent Technologies) The amplified fragments were purified and resuspended by using AMPure XP magnetic beads (Beckman Coulter, Brea, California) and Tris-HCL buffer (pH 8.0), respectively. Each sample library was kept at − 20 ◦C until sequencing.

The DNA libraries were quantified before sequencing using Qubit® 2.0 Fluorometer (Thermo Fisher Scientific) and 2100 Bioanalyzer (Agilent Technologies) were used for quantifying DNA libraries before sequencing. Pools of DNA libraries containing up to 96 samples (4 nmol/ L) were then diluted as recommended by the manufacturer. The sequencing of DNA libraries was performed by employing the MiSeq Reagent kit V3 (600 cycles), in the MiSeq Personal Sequencer (Illumina, San Diego, California).

2.3 Genotyping

Three software programs were used to perform genotype calling directly from FASTq files: HipSTR (Willems et al., 2017); STRait Razor (Warshauer et al., 2013); and toaSTR (Ganschow et al., 2018). The Integrative Genomics Viewer (IGV) software (Thorvaldsdo et al., 2013; Robinson et al., 2017) was also used to verify whether the target regions were successfully sequenced and to interpret discordant results.

For the HipSTR software the GitHub guide page steps were followed by using a 15% stutter model and 1 as analytical threshold (AT – minimum total coverage of reads) parameters. The VCF output file was organized in an Excel spreadsheet. The alleles were determined by using two parameters: the reference allele of each marker (available in the HipSTR repository) and the value of the base pair differences (GB) found in the VCF file. The VCF file also displays the total number of reads for each allele of each marker.

Table 1. Average coverage and standard deviations for the 22 STRs called by each genotyping software.

Markers	toaSTR		HipSTR		STRait Razor	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
CSF1PO	128.30	86.77	41.34	21.63	109.49	76.56
D1S1656	21.82	13.55	17.52	9.90	16.29	9.91
D2S441	295.03	185.54	63.43	29.20	245.98	163.92
D2S1338	16.45	10.88	19.35	9.51	13.57	8.67
D3S1358	120.46	96.22	67.40	35.91	83.62	60.53
D5S818	75.64	50.80	31.45	17.82	65.61	46.55
D7S820	51.51	37.79	26.44	15.12	34.81	28.14
D8S1179	54.73	45.46	19.58	10.12	49.16	36.53
D10S1248	46.52	28.16	26.10	13.54	18.19	13.42
D12S391	86.56	67.19	20.69	7.87	75.16	52.72
D13S317	134.84	87.79	37.86	20.23	96.44	70.60
D16S539	121.33	83.38	37.58	20.04	89.54	63.43
D18S51	58.77	42.55	19.26	9.17	44.70	34.84
D19S433	58.37	36.23	21.41	11.52	9.41	6.82
D21S11	36.82	23.05	62.12	35.11	12.13	7.24
D22S1045	66.34	48.60	20.29	11.50	63.20	47.26
FGA	32.37	23.62	33.84	20.71	21.70	17.44
Penta D	39.26	27.86	33.31	18.77	18.61	14.86
Penta E	12.97	7.59	12.73	5.82	-	-
TH01	426.04	321.13	44.88	25.22	150.55	110.15
TPOx	455.32	305.09	17.63	7.67	408.37	303.52
vWA	45.48	35.31	14.40	6.99	-	-
Mean	108.40	123.52	31.30	16.15	81.33	96.07

The analysis panel on the toaSTR platform was created by including the STR markers of interest. The following parameters were established: two mismatches (Mm – mutations, insertions, or substitutions present in the flanking regions), 2 as calling threshold (CT – minimum allele coverage), and 1 as the analytical threshold. The results were exported in PDF format and copied into Microsoft Excel.

STRait Razor v3 was run using the configFile available at: https://github.com/Ahhgust/STRaitRazor/blob/master/configFile. This file includes the necessary information to genotype all autosomal STRs included in the present study. As explained latter, flanking regions other than those present in this configFile were used to genotype five STRs: D2S441, D12S391, D22S1045, Penta E and vWA. The new sequences of such flanking regions are described in the Supplementary material (Supplementary Table 2). For sample processing with STRait Razor v3, the input files were loaded by using the Microsoft Excel plugin provided by the developers. The protocol described in the STRait Razor website was used, establishing AT = 2 and CT = 2 as parameters (Warshauer et al.,2013).

For all the three methods, sequencing coverages were directly provided by each program. Some rules were applied to correct possible interpretation errors, making the called genotype more reliable. Any single nucleotide mutation present within the repetitive regions were not take into account to determinate the nomenclature of the alleles. Thus, the nomenclature of the alleles was given only regarding their respective lengths. Homozygosis and heterozygosis were accepted with coverages of at least eight and four reads, respectively. This procedure ensured ($P > 0.99$) that a homozygous genotype is called because of a lack of variation at that position and not because the second allele was not sampled. Adjacent alleles with coverages below 15% were considered stutter. Genotypes that did not fit these rules were excluded. Finally, in order: (a) to identify incompatible genotypes and (b) to provide consensus genotypes for each locus in each sample, the genotypes determined by each program were submitted to an exhaustive comparison analysis, which was carried out by looking for identical genotypes, partial genotypes composed of one correct and one incorrect allele, partial genotype with a correct and doubtful allele, and totally discordant genotypes. As demonstrated in previous studies, the three programs have consistently presented high accuracy in genotype calling, as determined by different validation procedures (Willems et al., 2017; Warshauer et al., 2013; Ganschow et al., 2018). For this reason, an allele was considered correct when the methods agreed.

When the allele was suggested by only one program, it was deemed incorrect, disagreeing with any other allele found. Alleles that were determined by only two programs needed a more careful evaluation: when these alleles coincided, they were considered correct, but they were deemed doubtful when they differed. Finally, an allele was deemed doubtful when the three tools revealed different alleles (e.g., genotypes 7/7, 7/8, and 7/9 for HipSTR, STRait Razor, and toaSTR, respectively), resulting in three partial genotypes with incorrect alleles, although it is possible that one of these genotypes was in fact correct. From this consensus genotype calling procedure, the proportion of correct, incorrect, and doubtful alleles were also registered.

To investigate the influence of coverage in genotype calling and software comparisons, we have selected two subsets of samples. One group containing the 50 samples with the highest coverage and a second group containing the 50 samples with the lowest coverage. For selecting these samples, we considered average coverages provided by HipSTR, since it was the software that presented the lowest overall averages of reads (Table 1). These two groups ware also used to estimate the impact of coverage on heterozygote balance, considering the results from toaSTR and STRait Razor.

2.4 Statistical analysis

Allele frequencies and adherence to Hardy–Weinberg equilibrium expectations were calculated by using the software GenAlEx (Peakall et al., 2012). Forensic parameters, such as Match Probability (MP), Power of Discrimination (PD), Power of Exclusion (PE) and Polymorphic Information Content (PIC), were estimated by employing STRAF 1.0.5 (Gouy et al., 2017). Heterozygote balance across *loci* was compared by means of the Student’s paired t-test, using GraphPad InStat 3.06 (GraphPad Software, Inc).

3. RESULTS

The three programs were useful for genotyping STR markers captured with the HaloPlex Target Enrichment System kit, sequenced by the MiSeq Illumina platform. We included 547 samples in the study, but were disregarded 65 before STR genotype calling due to library preparation and sequencing issues that resulted in very low coverage and poor-quality reads.

3.1 Performance comparison

By using the IGV software (Thorvaldsdo etl al., 2013; Robinson et al., 2017) we verified that all the target regions were successfully sequenced. HipSTR and toaSTR genotyped the 22 markers, while STRait Razor completely failed to detect Penta E and had poor performance for calling vWA (only 1.24% of samples was called). HipSTR presented nomenclature problems in five markers: D19S433, D21S11, Penta D, Penta E, and vWA. These problems involved a shift of some base pairs in allele calling. We analyzed genotypes from these markers with IGV software (Thorvaldsdóttir et al., 2013; Robinson et al., 2017), HipSTR VizAln function (Willems et al., 2017), and we also compared their sequences with those from alleles recorded in STRBase. These three procedures led us to conclude that the problems in all five markers were related to the reference allele used by HipSTR to perform genotype calling. According to the nomenclature indicated by the ISFG (Gettings et al., 2019), we adjusted alleles from these makers as follows: we removed two repeat units from all D19S433, D21S11, and vWA alleles; we included one repeat unit in all Penta D alleles; and, for Penta E, we established the reference allele used for genotype calling was established as 5 instead of 5.2.

The average depth of coverage was different for each program. Overall, HipSTR presented the lowest coverage (31.30 ± 16.15, ranging from 12.73 ± 5.82 for Penta E to 67.40 ± 35.91 for D3S1358), whereas toaSTR retrieved the highest (108.40 ± 123.52, ranging from 12.97 ± 7.59 for Penta E to 455.32 ± 305.09 for TPOx). Table 1 shows the average coverage for each marker as assessed by the three programs.

Table 2. Summary of the possible causes of genotype disagreements among methods.

Software	**toaSTR**	**HipSTR**	**STRait Razor**	**Total**
Number of informative comparisons	9,586	8,801	7,969	26,356
Total number of disagreements	30 (0.3%)	141 (1.6%)	154 (1.9%)	325 (1.2%)
Stutter-related problems	12	81	18	111
Stutter-unrelated problems	18	60	136	214
Stutter-related problems				
Unable to detect the smallest adjacent allele (False homozygous)	3	8	14	25
Detection of a smaller nonexistent adjacent allele (False heterozygous)	9	73	4	86
Stutter-unrelated problems				
Detection of a nonexistent allele instead of a true allele in a heterozygous sample	9	30	4	43
Detection of a nonexistent non-adjacent allele in a homozygous sample (False heterozygous)	1	21	0	22
Unable to detect the largest allele (False homozygous)	8	6	74	88
Unable to detect the smallest allele (False homozygous)	0	3	58	61

Table3. Comparison of average genotype calling performance achieved with toaSTR, HipSTR and STRait Razor softwares from 482 Brazilian samples.

	toaSTR	HipSTR	STRait Razor[a]
Average number of genotypes called	454±42	410±39	399±106
Average coverage	112±168	32±24	93±138
Average number of informative genotypes[b]	436	400	398
Genotypes			
Identical genotypes[c]	96.13%	94.01%	95.73%
Partial genotype with a correct and doubtful allele	3.07%	2.60%	0.92%
Partial genotypes composed of one correct and one incorrect allele	0.79%	3.29%	3.34%
Totally discordant genotypes	0.01%	0.01%	0.01%
Alleles			
Correct alleles	98.06%	97.05%	97.86%
Doubtful alleles	1.54%	1.30%	0.46%
Incorrect alleles	0.40%	1.66%	1.68%

[a] The averages estimated for STRait Razor do not include the two STRs (vWA and Penta E) that were not genotyped by it.

[b] Genotypes called by the given method and at least another method.

[c] Percentage estimated from the number of informative genotypes presented above.

Two samples showed an intriguing inconsistency in D18S51 when we compared results from the three methods: different heterozygous genotypes from the combination of three different alleles. After careful analysis and visualization of sequencing reads, we confirmed triallelic patterns in both samples (14, 15, and 17 in one of them and 10, 12, and 14 in another).

Besides this situation, which emphasized the difficulty in dealing with tri-allelic genotypes, we observed non-biological discrepancies in the comparison of results, which must be addressed. Supplementary Table 3 details the possible causes of disagreements between the genotypes determined by each program for each STR. Table 2 presents a summary of these disagreements. We classified them into two groups: those related to confusion usually caused by stutters and those concerning other causes. STRait Razor showed a larger number of differences (154), from which 74 (48%) consisted in the failure in capturing larger alleles. HipSTR presented 141 disagreements, from which 81 (57.4%) were due to stutter. Finally, toaSTR presented a much smaller number of divergences, and they are evenly distributed across three classes of disagreements (Table 2); notwithstanding that, 19 (63.3%) out of the 30 disagreements involve the detection of a non-existent allele.

Table 3 compares the average performance of the three methods. toaSTR was capable of retrieving a larger number of genotypes than the other two methods. Although STRait Razor presented a much lower genotype calling rate (82.78%), accuracy levels for genotype calling were similar (identical genotypes ~ 95%, partial genotypes ~ 5%, and totally discordant genotypes ~ 0.01%) across the three methods. The major difference lay in the proportion of partial genotypes composed of one correct allele and one doubtful or incorrect allele: most of toaSTR and STRait Razor partial genotypes were composed of doubtful and incorrect alleles, respectively. The concordance index was even higher for alleles than for genotypes. Likewise, accuracy levels for allele calls were very similar (correct alleles ~ 97.5%) across the three methods, with toaSTR and STRait Razor presenting larger proportions of doubtful and incorrect alleles, respectively.

Supplementary Table 4 details how each program performs for each marker. We observed that the worst performance concerned Penta E. Besides not being retrieved by STRait Razor, it presented the lowest average proportions of correct alleles (92.6%) and identical genotypes (85.3%), coupled with the smallest number of genotypes called (251) by toaSTR and HipSTR. In general, we verified the best performances for D3S1358 and CSF1PO, with the highest average proportions of correct alleles (99.7%, and 99.5%, respectively) and identical genotypes (99.4% and 99.1%, respectively), coupled with the largest averages of genotypes called (466.00 and 467.33, respectively). Supplementary Figure 1 shows Venn diagrams addressing the concordance rates between the different programs. Considering only those samples that had their genotypes called by all three software for a given marker, D3S1358 showed the highest concordance between the genotypes obtained (433 of 442, i.e., 97.96%), while FGA showed the lowest (248 of 307, i.e., 80.78%) (Fig. 1).

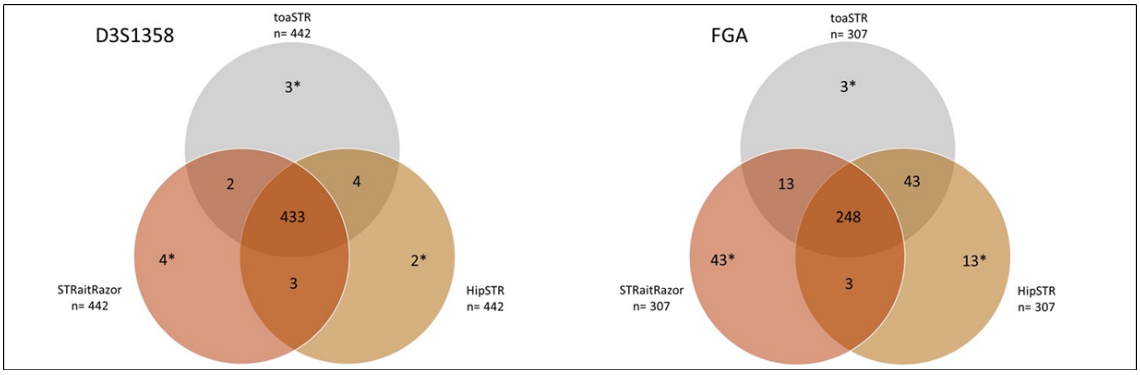

Figure 1. Venn diagrams of D3S1358 and FGA, constructed from those samples that had their genotypes called by the three software programs. Asterisks indicate the number of discordant genotypes for each program.

Supplementary Tables 5 and 6 depicts the influence of coverage in genotyping efficiency. In the lowest-coverage subset, 24.5% of the genotypes (269 out of 1100) were discarded, 41 of them due to inconsistencies between the three software. On the other hand, in the highest-coverage group, only 2.5% of the genotypes (28 out of 1100) were discarded, 10 of which due to inconsistencies. The other discarded genotypes were not considered because it was not possible to genotype them in two or more software. The group of 50 samples with the highest coverage presented six *loci* (CSF1PO, D3S1358, D5S818, D7S820, D10S1248, and D13S217) with fully concordant genotypes among all three programs, and eight *loci* (D1S1656, D2S441, D8S1179, D16S539, D18S51, Penta D, TH01, and TPOX) with fully concordant genotypes in two out of the three software. Conversely, when considering the lowest-coverage subset, only D19S433 presented fully concordant genotypes among all three programs, and only two (D5S818 and Penta D) in two out of the three software (Supplementary Table 5). Surprisingly, the three software presented increased rates of identical genotypes for D19S433 in the 50 lowest-coverage samples, despite the reduced numbers of informative comparisons. The same issue happened with D2S441, D21S11, D22S1045, and FGA concerning STRait Razor. When considering the averages across *loci* (Supplementary Table 6), the conclusion that higher coverage results in increased efficiency and accuracy is straightforward: no totally discordant genotypes was observed for the subset of highest coverage and the proportion of correct alleles called ranged from 97.78% (STRait Razor) to 99.30% (toaSTR).

Interestingly, STRait Razor was the software that presented the highest proportions of identical genotypes (94.22%) and correct alleles (97.11) in the subset of samples with lowest coverage, but it was also the soft- ware that resulted in the lowest number of informative comparisons in both groups, 45.65 (ranging from 16 to 50) in the highest-coverage subset and 32.41 (ranging from 7 to 49) in the lowest-coverage.

These two groups ware also used to estimate the impact of coverage on heterozygote balance (Supplementary Table 7). This analysis was performed considering only the results from toaSTR and STRait Razor, since HipSTR does not provide the number of reads of each allele from a given genotype. In general, the average ratio of reads observed between the allele with less reads and the allele with more reads in a given genotype was approximately 0.7. While the coverage did not have a significant impact on heterozygote balance from genotypes determined by STRait Razor (p = 0.1182), it had a significant impact on genotypes from toaSTR (p = 0.0002). Interestingly, considering the 50 lowest coverage samples, mean heterozygote balance did not differ between the two software (p = 0.0928), but differed significantly for the 50 highest-coverage samples (p = 0.0005). Overall, when coverage is high, toaSTR is capable of reducing the difference in coverage between the alleles with less and more reads in a given heterozygote.

3.2 Allele frequencies and forensic parameters

Table 4 lists the allele frequencies and forensic parameters calculated for the final genotypes (consensus among the three programs). Final genotypes do not include genotypes determined by only one program or with doubtful alleles. However, the results of independent analysis considering all the genotypes determined by each program are available (Supplementary Tables 8–10).

Table 4. Forensic parameters and allele frequencies for the 22 STRs analyzed in the Brazilian population sample.

Allele/n	CSF1PO	D1S1656	D2S441	D2S1338	D3S1358	D5S818	D7S820	D8S1179	D10S1248	D12S391	D13S317	D16S539	D18S51	D19S433	D21S11	D22S1045	FGA	Penta D	Penta E	TH01	TPOX	vWA
2.2																		0.045				
3.2																		0.001				
5																		0.019	0.096		0.001	
6	0.001																	0.001		0.209	0.018	
7	0.016					0.011	0.009					0.001						0.011	0.175	0.238	0.004	
8	0.017		0.002			0.018	0.174	0.006			0.117	0.024						0.045	0.072	0.125	0.459	
8.3																				0.001		
9	0.033		0.003			0.038	0.124	0.016	0.001		0.086	0.156	0.002	0.001		0.001		0.192	0.019	0.188	0.138	
9.3																				0.230		
10	0.265	0.006	0.210			0.057	0.276	0.078	0.003		0.056	0.092	0.007	0.006		0.017		0.134	0.058	0.010	0.065	
10.1			0.001																			
10.3	0.001		0.001																			
11	0.317	0.052	0.317		0.001	0.316	0.233	0.075	0.009		0.322	0.313	0.013	0.025		0.115		0.164	0.107		0.269	0.003
11.3			0.051																			
12	0.289	0.110	0.069			0.362	0.149	0.147	0.065		0.256	0.236	0.142	0.098		0.013		0.149	0.140		0.045	
12.2														0.012								
12.3			0.002																			
13	0.055	0.075	0.029		0.003	0.186	0.031	0.258	0.257		0.122	0.155	0.126	0.252		0.005		0.160	0.107		0.001	0.005
13.1			0.001																			
13.2													0.001	0.030								
13.3			0.001																			
14	0.006	0.153	0.266		0.100	0.012	0.004	0.241	0.327		0.039	0.018	0.139	0.288		0.060		0.057	0.075			0.089
14.2													0.001	0.043								
14.3		0.007																				
15		0.167	0.043	0.003	0.267	0.001		0.137	0.192	0.054	0.002	0.005	0.125	0.140		0.349		0.020	0.061			0.152
15.2														0.051								
15.3		0.041																				
16		0.106	0.001	0.066	0.269			0.031	0.119	0.023			0.137	0.031		0.301		0.002	0.044			0.230
16.1																	0.001					
16.2														0.017								
16.3		0.057																				
17		0.044	0.001	0.231	0.204			0.009	0.023	0.108			0.128	0.002		0.132	0.001		0.023			0.255
17.1										0.001												
17.2														0.005								
17.3		0.116								0.013												
18		0.005		0.084	0.145			0.001	0.003	0.196			0.071			0.007	0.013		0.002			0.180
18.1										0.001												
18.2																	0.006					
18.3		0.050								0.020												
19				0.113	0.009					0.159			0.059				0.072		0.019			0.069
19.1										0.001												
19.2																	0.001					
19.3		0.007								0.006												
20				0.133	0.001					0.122			0.035				0.110					0.017
20.1										0.001												

20.2													0.001				0.002					
20.3		0.004								0.002							0.001					
21				0.034						0.096			0.009				0.153					0.002
21.2																	0.002					
22				0.066						0.096			0.003				0.159					
22.2																	0.006					
22.3																	0.001					
23				0.121						0.059							0.160					
24				0.069						0.030			0.002				0.142					
24.2															0.004							
25				0.063						0.010			0.001				0.101					
25.2															0.001		0.001					
26				0.012						0.001					0.001		0.036					
27				0.006											0.036		0.017					
27.1																	0.001					
28															0.152		0.009					
28.3															0.003							
29															0.237		0.001					
30															0.232							
30.2															0.029							
31															0.068							
31.2															0.098		0.001					
32															0.006							
32.2															0.078							
33.2															0.039							
34															0.005							
34.2															0.004							
35															0.005							
35.1															0.001							
N	478	406	468	335	478	469	459	466	448	471	471	472	461	422	398	410	412	403	214	477	476	320
Na	10	16	16	13	9	9	8	11	10	20	8	9	19	15	18	10	24	14	14	7	9	10
Ho	0.707	0.936	0.786	0.842	0.828	0.695	0.830	0.805	0.777	0.870	0.790	0.788	0.902	0.832	0.842	0.763	0.864	0.836	0.911	0.767	0.687	0.834
He	0.741	0.894	0.774	0.876	0.784	0.729	0.801	0.822	0.771	0.883	0.790	0.789	0.884	0.817	0.842	0.753	0.876	0.862	0.896	0.796	0.691	0.814
MP	0.114	0.025	0.089	0.031	0.086	0.113	0.071	0.056	0.093	0.025	0.071	0.079	0.026	0.057	0.042	0.113	0.029	0.040	0.021	0.073	0.137	0.063
PD	0.886	0.975	0.911	0.969	0.914	0.887	0.929	0.944	0.907	0.975	0.929	0.921	0.974	0.943	0.958	0.887	0.971	0.960	0.979	0.927	0.863	0.937
PIC	0.694	0.887	0.744	0.871	0.751	0.687	0.775	0.802	0.734	0.876	0.766	0.759	0.876	0.796	0.827	0.717	0.865	0.838	0.891	0.766	0.658	0.785
PE	0.460	0.887	0.610	0.850	0.652	0.454	0.657	0.645	0.602	0.761	0.616	0.599	0.833	0.648	0.687	0.599	0.748	0.660	0.738	0.555	0.463	0.624

N: number of individuals analyzed; Na: number of alleles found; Ho: Observed heterozygosity; He: Expected heterozygosity; MP: Match probability; PD: Discrimination power; PIC: Polymorphic information content; PE: Exclusion power

PE values ranged from 0.408 (TPOX) to 0.818 (D21S11 and Penta E), and PM values varied from 0.025 (D21S11 and Penta E) to 0.140 (TPOX). Combined match probability was 8.17×10−26 for STRait Razor, 4.09×10−28 for HipSTR, 3.70×10−28 for toaSTR, and 2.90×10−28 for the final genotypes. The combined exclusion power obtained from STRait Razor was 0.99999997355, from HipSTR 0.99999999998, from toaSTR 0.99999999948 and from the final genotypes 0.99999999982. The absence of vWA and Penta E largely influenced the poorer performance of STRait Razor.

All the markers were in Hardy–Weinberg equilibrium when final genotypes were considered. Considering the results of each program individually, STRait Razor showed departures from equilibrium in six markers (D2S1338, D19S433, D21S11, D22S1045, FGA, and Penta D), HipSTR in two (D2S1338 and D22S1045) and toaSTR in one (D2S1338) (Table 5).

Table 5. Probability of adherence to Hardy–Weinberg Equilibrium expectations (*p*HWE) for the 22 STR markers genotyped by STRait Razor, HipSTR, and toaSTR, as well as for final (consensus) genotypes from the Brazilian population sample.

Locus	STRait Razor (*p*HWE*)*	HipSTR (*p*HWE)	toaSTR (*p*HWE)	Final genotypes (*p*HWE)
CSF1PO	0.9032	0.9567	0.7137	0.8813
D1S1656	0.2063	0.3180	0.2009	0.0916
D2S1338	**0.0025**	**0.0410**	**0.0042**	0.4345
D2S441	**0.0001**	0.5913	0.8326	0.8558
D3S1358	0.6259	0.5515	0.4390	0.5522
D5S818	0.8802	0.9962	0.8650	0.8307
D7S820	0.2315	0.6149	0.8788	0.8853
D8S1179	0.1947	0.8548	0.5627	0.5330
D10S1248	0.9751	0.8556	0.9660	0.9020
D12S391	0.9455	0.9531	0.9436	0.9301
D13S317	0.6880	0.4346	0.5943	0.5965
D16S539	0.6691	0.6223	0.5115	0.5853
D18S51	0.9756	0.7145	0.9061	0.9494
D19S433	**0.0000**	0.9995	0.9886	0.9981
D21S11	**0.0000**	0.9999	10.000	10.000
D22S1045	0.1757	**0.0078**	0.0672	0.2806
FGA	**0.0000**	10.000	0.7590	10.000
TH01	0.1785	0.7378	0.1780	0.1967
TPOx	0.5666	0.4313	0.5760	0.5695
vWA	-	0.8817	0.1619	0.9565
Penta D	**0.0003**	0.1944	0.0728	0.1841
Penta E	-	0.3339	0.0895	0.3181

**Bold values indicate significant deviation from equilibrium.*

4. DISCUSSION

This study compares the STR genotypes called by three computational methods, in 482 Brazilian individuals evaluated with NGS. The three programs have proven useful for genotyping STRs, but STRait Razor completely failed in genotyping vWA and Penta E. However, these markers have been previously genotyped by this tool in other studies, which suggests that this issue may be specific for the sequencing assay used in the present study.

For proper calling of the *loci* of interest that the flanking regions must be completely sequenced in high quality. All three software works with the raw sequence data available in the FastQ files and do not require prior trimming and cleanup. However, we manually checked the quality of alignments using the IGV software, seeking to identify regions with an excessive amount of bases sequenced with quality lower than Q30, which would impair recognition of the reads by the software. Each software uses different segments of the flanking regions, different types of algorithms to perform genotype calls, and different approaches to estimate coverage, which explains why the methods vary in terms of coverage and genotype calling efficiency. One of the most important differences is that while toaSTR and STRait Razor use flanking regions to capture the STR markers, HipSTR is based on a hidden Markov model (HMM) to realign the STR-containing reads to candidate haplotypes that eventually include SNPs, using, therefore, the whole STR genomic location (including). More information on the genomic regions are available at Supplementary Table 11.

Here, the regions that the developers of STRait Razor suggest for genotyping D2S441, D12S391, D22S1045, vWA, and Penta E have not been fully covered in the sequencing or, when sequenced, present low quality, as confirmed by direct analysis of the reads with the IGV software. MiSeq Illumina sequencing is characterized by decreased quality of the bases incorporated in the last sequencing cycles. In this way, STRs with larger amplicons may have their flanking region sequenced with low quality, making bioinformatics analysis difficult (Bornman et al., 2012). The paired-end reads generated by the MiSeq Reagent kit V3 sequencing used in this study presents a maximum length of 300 bases. This length, added to a lower sequencing quality in the latest bases incorporated into reads, creates difficulties in obtaining long alleles from STRs that require longer reads that fully encompassing their flanking regions (Erlich et el., 2008; Yang et al., 2013). In the present case, this could interfere in read alignment and STR calling by each program in varying levels. Thus, we have identified new flanking regions closer to the repetitive regions and used them to genotype these five markers (Supplementary Table 2). Even with these new flanking regions cannot provide good results for Penta E and vWA. We have detected low coverage vWA alleles in several samples, while we have obtained no results for Penta E. On the other hand, we have recovered D2S441, D12S391, and D22S1045 genotypes in 474, 469, and 390 samples, respectively.

In fact, coverage was highly variable, particularly for toaSTR and STRait Razor (Table 1). At first, we thought that the number of different probes designed through SureDesign to capture each STR (Supplementary Table 1), which range from 2 to 7 (mean = 4.27 ± 1.39), could be responsible for such variation. However, although it may be influencing the low coverage for some STRs (e.g., D1S1656 and vWA), it does not explain the low coverage for others (e.g., D2S1338 and Penta E). Moreover, there is no statistical correlation between the number of designed probes for each STR and coverage obtained with any of the three software. Therefore, other factors may be influencing coverage as well. As reported in other studies, difficulties in genotyping STRs and alleles that require larger amplicons are expected (Fungtammasan et al., 2015), as explained above. The average coverages of larger STRs, such as D2S1338, D19S433, FGA, Penta D, and Penta E, are low when compared to shorter markers, such as D2S441, TH01, and TPOx (Table 1). In fact, average coverage was negatively correlated with size of the repetitive sequence for toaSTR ($r = -0.7293$; $p = 0.0001$) and STRait Razor ($r = -0.6528$; $p = 0.0018$), but not for HipSTR ($r = -0.2735$; $p = 0.2181$), which may be reflecting the fact that HipSTR considers the whole STR genome sequence rather than only the flanking region for genotype calling (Supplementary Table 11). This clearly reflects the coverage obtained for TPOx, which presented the highest coverage in both toaSTR and STRait Razor, and a very low coverage for HipSTR. Notwithstanding, it should be kept in mind that if the number of samples in each sequencing assay is reduced, a higher coverage will be obtained. Overall, the coverage pattern described in this study resembles the one observed by Wendt and collaborators (Wendt et al., 2016) by using STRaitRazor and Haloplex/MiSeq.

The number of samples with called genotypes, the proportion of identical genotypes among methods, and the correct genotype rate (Supplementary Table 4) suggest that toaSTR performs better for D3S1358, D7S820, D8S1179, D10S1248, D16S539, D18S51, D19S433, and TPOx; HipSTR for D21S11, and D22S1045; and STRait Razor for CSF1PO, D5S818 and TH01; however, their performance for each marker is very similar. For five markers, HipSTR (D1S1656, D13S317, FGA, and Penta D) and STRait Razor (D2S1338) present a better accuracy than toaSTR at the expense of genotyping a smaller number of samples. Finally, because only toaSTR and HipSTR can successfully genotype vWA and Penta E, the accuracy parameters of both tools are identical. The rate of disagreements is clearly influenced by coverage. Those samples with lower coverage are more impacted by the presence of stutter or other artifacts that generate incorrect genotypes. As expected, Supplementary Tables 5 and 6 demonstrate that samples with higher coverage present higher efficiency in genotype calling and concordance in the genotypes and alleles determined by the three programs. Genotype calling should be performed by using more than one software. Despite the problems described above, mainly related to the sequencing technology and not to the software accuracy, the three programs have provided good results. However, the choice may rely upon careful evaluation of the advantages and disadvantages of each software. toaSTR analyzes one sample per time, requiring more time and manual work. Therefore, it is useful to process a few samples. It has a graphical interface with interactive visualization of each marker, which assists analysis of mixtures, definition of doubtful genotypes, and identification of stutters. Besides, it is simple, and dismisses the need for prior knowledge in bioinformatics. Unlike toaSTR, HipSTR and STRait Razor can process hundreds of samples at once. They allow addition of new markers or modification of the flanking regions to capture them (Willems et al., 2017; Warshauer et al., 2013). The output of HipSTR, as already described, does not provide the alleles directly, so additional actions are necessary to calculate them by using the resulting VCF, reference alleles and parameters defined by the program. Moreover, processing with HipSTR is more complex and re- quires some knowledge in bioinformatics.

HipSTR presents nomenclature problems for D19S433, D21S11, Penta D, Penta E and vWA. The genotypes calculated by using HipSTR differ from those obtained with toaSTR and STRait Razor, involving a clear pattern of shift of some base pairs in all called alleles. We have compared the results from each program by using the IGV software This confirmed the nomenclature problem, so the HipSTR results had to be adjusted to be in accordance with the ISFG nomenclature (Gettings et al., 2019) followed by the other two methods. Regarding coverage, in contrast to toaSTR and STRait Razor, HipSTR, does not provide separate coverages for each allele (only full coverage). Therefore, we have not been able to apply some of the coverage rules established in the Materials and Methods section to assess the reliability of each allele in the calculated genotypes.

It should be mentioned that most of the forensic labs employ amplicon-based rather than hybridization capture assays. However, this three software were developed to present accurate performances with NGS data produced by different sequencing platforms, either using or not commercially available STR kits, such as Precision ID Global-FilerTM NGS STR Panel v2 (Thermo Fisher) and Forenseq DNA Signature Prep kit (Illumina). These three programs present high consistency levels when their results are compared with capillary electrophoresis (CE). HipSTR accuracy was tested by comparing whole-genome sequencing calls from 118 samples from the Simons Genome Diversity Project (Mallick et al., 2016), sequenced with a 100-bp paired-end PCR-free protocol (Illumina) to capillary electrophoresis CE data, which gave 98.8% consistency (Willems et al., 2013). Ganschow et al. (2018), using different commercial kits, such as Nextera XT (Illumina), obtained 171 length- and sequenced-based genotypes with toaSTR that were 100% concordant with CE. Warshauer et al. (2013) described similar results: allele calls made by STRait Razor using both the TruSeq Custom Enrichment protocol (Illumina) and the HaloPlex Target Enrichment protocol (Agilent Technologies), were fully concordant with the CE genotypes (Warshauer et al., 2013).

Here, we have compared the genotypes called by each software to analyze the discordant genotypes better. Overall, we verified no new alleles were observed. However, during the analysis, two samples showed inconsistencies in D18S51 genotypes due to triallelic patterns in both samples. Chromosomal trisomy and localized duplications or length mutations during DNA replication and cell division can cause triallelic patterns (Huel et al., 2007; Mertens et al., 2009). According to the

STRbase (https://strbase. nist.gov/var_D18S51.htm#Tri) and other studies (Peakall et al., 2012; Gouy et al., 2017), D18S51 triallelic genotypes have been previously observed.

Here, the disagreements are predominantly stutter-unrelated problems. Notwithstanding, different problems affect the three methods. By analyzing the disagreements in the genotypes determined by STRait Razor, we have observed 146 false homozygous calls (Table 2), which correspond to 94.81% of the erroneous calls. More than 39% of these errors affect the D2S441 marker (52 samples), particularly in situations involving non-consensus alleles, such as 11.3 and 12.3. On the other hand, 94 out of 141 HipSTRs disagreements (66.67%) involves identification of non-existent alleles in homozygous samples (false heterozygotes), mostly due to stutter. Thirty-four out of the 141 disagreements found in HipSTR are related to the D22S1045 marker. These problems affect mainly alleles with low coverage (between 8 and 20). Finally, toaSTR provides only 30 disagreements, which represents a fivefold smaller rate.

These differences are reflected in the adherence to the Hardy–Weinberg equilibrium (HWE) proportions. With STRait Razor results, six markers have significant deviations from HWE: D2S1338 (*p* = 0.0025), D2S441 (*p* = 0.0001), D19S433 (*p* = 0.0000), D21S11 (*p* = 0.0000), FGA (*p* = 0.0000), and Penta D (*p* = 0.0003). Although we have observed deficit of heterozygosis in all situations – D2S1338 (11%), D2S441 (8%), D19S433 (2%), D21S11 (6%), FGA (13%), and Penta D (8%) – the deviation involving D19S433 is possibly a consequence of the high number of samples (337, i.e., 69.92% of the samples) not called, which usually involves longer alleles. HipSTR shows two significant deviations: D2S1338 (*p* = 0.0410) and D22S1045 (*p* = 0.0078). In toaSTR data, however, only one marker is in disequilibrium: D2S1338 (*p* = 0.0042), with 8% heterozygosity deficiency. Due to the significance level adopted herein (α = 0.05), approximately 5% (1.1 in 22 markers) of random deviations are expected. Nevertheless, the dataset composed of the final genotypes (consensus) does not disclose significant deviations in relation to the Hardy–Weinberg equilibrium, supporting the hypothesis that genotype calling involving more than one type of software minimizes errors, resulting in more reliable and accurate outcomes.

Comparison of the allele frequencies from the present study with those from other Brazilian population samples analyzed by CE (Moysés et al., 2017; Aguiar et al., 2012), does not show any substantial differences. However, here we have not found the smallest and longest alleles from every locus (Powerplex 16) previously observed in a very large Brazilian population sample (*n* = 137,161), in which these alleles usually presented frequencies much lower than 0.1%. An opposite scenario arises when a smaller (*n* = 208) population sample from Southeastern Brazil is taken into ac- count: 22 rare alleles (21 of them with a single copy and one with two copies) from the Globalfiler set of STRs have not been sampled herein, whilst that study missed 33 alleles we observed in this study. Anyway, given the low frequency of these alleles in the other Brazilian populations, their absence in the present sample is probably due to chance rather genotype calling errors.

As expected, the forensic parameters indicate a highly informative set of markers, with values resembling those of other studies with worldwide populations (Table 6) (Butler et al., 2012). It is important to note that the forensic parameters estimated by using the genotypes obtained by STRait Razor are based on 20 of the 22 *loci*, which explains the lower match probability (8.17 × 10^{-26}) and exclusion power (0,9999999973553).

Table 6. Match probability and exclusion power estimates obtained in worldwide populations.

Population sample	*N*	Number of STRs	Genotyping strategy	Match probability	Exclusion power	Reference
Brazil (Ribeirão Preto)	482	22	toaSTR (NGS)	3,70 x 10^{-28}	0.99999999948	This study
Brazil (Ribeirão Preto)	482	20	STRait Razor (NGS)	8,17 x 10^{-26}	0.99999997355	This study
Brazil (Ribeirão Preto)	482	22	HipSTR (NGS)	4,09 x 10^{-28}	0.99999999998	This study
Brazil (Ribeirão Preto)	482	22	All softwares - consensus (NGS)	2,90 x 10^{-28}	0.99999999982	This study
Brazil (Southeast)	208	21	GlobalFiler® (CE)	7.40 x 10^{-27}	0.9999999993	Moysés et al., 2017
Brazil (all regions)	137,000	15	PowerPlex®16 (CE)	2.45 x 10^{-18}	0.9999990272	Aguiar et al., 2012
Caucasians, African Americans, Hispanics and Asians	1,036	20	CODIS	9,35 x 10^{-24}	-	Butler, 2012

5. CONCLUSION

Despite locus-specific differences and better performance of toaSTR genotypes determined by HipSTR, STRait Razor, and toaSTR from NGS data in the Brazilian population are highly concordant and reliable. In the present study, the three software provided consistent STRs genotyping; however, we recommend using more than one software especially in cases of low coverage. Sequencing chemistries (Illumina, Thermo, Oxford Nanopore, and PacBio), equipment, and library preparation strategies (amplicon-based or hybridization capture), as well as random variables concerning the execution of library and sequencing protocols, may result in sequencing data that range from inadequate to optimum coverage and quality. Given that, other studies similar to this one is required in order to evaluate the performance of such tools and to indicate whether or not multiple software are recommended to enhance accuracy and reliability of called genotypes in various scenarios. However, the use of multiple software, even when not clearly required, will certainly assist with the identification of triallelic genotypes.

Notwithstanding, the complexity regarding STR genotyping from massive parallel sequencing data must be highlighted. Therefore, additional efforts are still necessary to improve the genotype calling accuracy from different next generation sequencing platforms.

6. Acknowledgments

Special thanks to Dr. Thomas Willems and Dr. Jonathan King for their support with the HipSTR and STRait Razor tools. We also thank Dra. Cynthia Maria de Campos Prado Manso for language editing. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. C.T.M.J. (#312802/2018-8) is supported by a Research fellowship from CNPq/ Brazil. This study was supported by FAPESP/Brazil (Fundação de Amparo à Pesquisa do Estado de São Paulo) Grants #2013/15447-0.

7. Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2022.102676.

8. REFERENCES

Aguiar V.R.; Wolfgramm E; Malta F.S.; Bosque A.G., et al. (2013). Updated Brazilian STR allele frequency data using over 100,000 individuals: an analysis of CSF1PO, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, FGA, Penta D, Penta E, TH01, TPOX and vWA loci. Forensic Sci Int Genet. 6(4):504-9. Doi: 10.1016/j.fsigen.2011.07.005.

Berglund E.C.; Lindqvist C.M.; Hayat S; Overnas E,. et al (2013). Accurate detection of subclonal single nucleotide variants in whole genome amplified and pooled cancer samples using HaloPlex target enrichment, BMC Genom. 14, 856. Doi: 10.1186/1471-2164-14-856

Bornman, D.M.; Hester, M.E.; Schuetter, J.M.; Kasoji, M.D., et al (2012). Short-read, high-throughput sequencing technology for STR genotyping. Biotech. Rapid Dispatches 1–6. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4301848/

Butler J.M.; Carolyn R.; Coble H.M.D (2012). Variability of New STR Loci and Kits in US Population Groups, 2012, pp. 1–28. Available from: https://strbase.nist.gov/pub_pres /Profiles-in-DNA_Variability-of-New-STR-Loci.pdf.

Chen M.C.; Sio C.P.; Lu Y.L.; Chang H.T., et al (2014). Identification of conserved and polymorphic STRs for personal genomes. BMC Genom. 15 (Suppl. 10), S3. Doi: 10.1186/1471-2164-15-S10-S3

Durney B.C.; Crihfield C.L.; Holland L.A. Capillary electrophoresis applied to DNA: determining and harnessing sequence and structure to advance bioanalyses (2009- 2014). Anal. Bioanal. Chem. 407 (23):6923–6938. Doi: 10.1007/s00216-015-8703-5.

Erlich Y; Mitra P.P.; de la Bastide M; McCombie W.R., et al (2008). Alta-cyclic: a self-optimizing base caller for next-generation sequencing, Nat. Meth 5(8):679–682.Doi: 10.1038/nmeth.1230

Fan H.; Chu J.Y (2007). A brief review of short tandem repeat mutation. Genom. Proteom. Bioinform. 5(1):7–14. Doi: 10.1016/S1672-0229(07)60009-6

Fungtammasan, A.; Ananda, G.; Hile, S.E.; Su, M.S., et al (2015). Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. Genome Res. 25, 736–749. Doi: 10.1101/gr.185892.114.

Ganschow, S.; Silvery, J.; Kalinowski, J.; Tiemann, C (2018). toaSTR: A web application for forensic STR genotyping by massively parallel sequencing. Forensic Sci. Int. Genet. 37, 21–28. Doi: 10.1016/j.fsigen.2018.07.006.

Gettings, K.B.; Ballard, D.; Bodner, M.; Borsuk, L.A., et al (2019). Report from the STRAND Working Group on the 2019 STR sequence nomenclature meeting. Forensic Sci. Int. Genet. 43, 102165. Doi: 10.1016/j.fsigen.2019.102165.

Gouy, A.; Zieger, M (2017). STRAF-A convenient online tool for STR data evaluation in forensic genetics. Forensic Sci. Int. Genet. 30, 148–151. Doi: 10.1016/j.fsigen.2017.07.007.

Gymrek, M.; Golan, D.; Rosset, S.; Erlich, Y (2012). lobSTR: A short tandem repeat profiler for personal genomes. Genome Res. 22, 1154–1162. Doi: 10.1101/gr.135780.111.

Hendricks K; McGuigan J; Snyder-Leiby T; Wiegand M., et al (2017). GeneMarker®HTS (High Throughput Sequencing) mtDNA Analysis Software for Next Generation Sequencing Data. Available from: https://www.bioke.com/blobs/manuals/SG/GMHTS_2017_AppNote.pdf

Huel RL, Basić L, Madacki-Todorović K, Smajlović L, et al., (2007). Variant alleles, triallelic patterns, and point mutations observed in nuclear short tandem repeat typing of populations in Bosnia and Serbia. Croat Med J. 48(4):494-502. Available at: https://pubmed.ncbi.nlm.nih.gov/17696304/

Koboldt D.C.; Steinberg K.M.; Larson D.E.; Wilson R.K., et al (2013). The next-generation sequencing revolution and its impact on genomics. Cell 155(1): 27–38. Doi: 10.1016/j.cell.2013.09.006.

Mallick, S., Li, H., Lipson, M. et al (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 538, 201–206 Doi: 10.1038/nature18964

Mertens G; Rand S; Jehaes E; Mommers N., et al (2009). Observation of tri-allelic patterns in autosomal STRs during routine casework, Forensic Sci. Int. Genet. Suppl. Ser. 2: 38–40. Doi: 10.1016/j.fsigss.2009.07.005

Miller S.A; Dykes D.D; Polesky H.F (1988). A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res.16(3):1215. Doi: 10.1093/nar/16.3.1215

Moysés C.B.; Tsutsumida W.M.; Raimann P.E.; da Motta C.H., et al. (2017). Population data of the 21 autosomal STRs included in the GlobalFiler® kits in population samples from five Brazilian regions. Forensic Sci Int Genet 26:28-30. Doi: 10.1016/j.fsigen.2016.10.017.

Peakall, R.; Smouse, P.E (2012). GenAlEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research-an update. Bioinformatics 28, 2537–2539. Doi: 10.1093/bioinformatics/bts460

Qin D (2019). Next-generation sequencing and its clinical application. Cancer Biol. Med. 16(1):4–10. Doi: 10.20892/j.issn.2095-3941.2018.0055

Robinson, J.T.; Thorvaldsdóttir, H.; Wenger, A.M.; Zehir, A., et al (2017). Variant Review with the Integrative Genomics Viewer. Cancer Res. 77, e31–e34. Doi: 10.1158/0008-5472.CAN-17-0337

Samorodnitsky E; Jewell B.M; Hagopian R; Miya J., et al. (2015). Evaluation of hybridization capture versus amplicon-based methods for whole- exome sequencing, Hum. Mutat. 36(9):903–914. Doi:10.1002/humu.22825

Tang H; Nzabarushimana E (2017). STRScan: targeted profiling of short tandem repeats in whole-genome sequencing data, BMC Bioinform.18(Suppl. 11): 398. Doi: 10.1186/s12859-017-1800-z.

Thorvaldsdóttir, H.; Robinson, J.T.; Mesirov, J.P (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief. Bioinform. 14, 178–192. Doi: 10.1093/bib/bbs017.

Warshauer, D.H.; Lin, D.; Hari, K., et al (2013). STRait Razor: A length-based forensic STR allele-calling tool for use with second generation sequencing data. Forensic Sci. Int. Genet. 7, 409–417. Doi: 10.1016/j.fsigen.2013.04.005

Wendt FR, Zeng X, Churchill JD, King JL., et al (2016). Analysis of Short Tandem Repeat and Single Nucleotide Polymorphism Loci From Single-Source Samples Using a Custom HaloPlex Target Enrichment System Panel. Am J Forensic Med Pathol. 37(2):99-107. Doi: 10.1097/PAF.0000000000000228.

Willems, T.; Gymrek, M.; Highnam, G.; Mittelman, D., et al (2014). The landscape of human STR variation. Genome Res. 24, 1894–1904. Doi: 10.1101/gr.177774.114.

Willems, T.; Zielinski, D.; Yuan, J.; Gordon, A., et al (2017). Genome-wide profiling of heritable and de novo STR variations. Nat. Methods 14, 590–592. Doi: 10.1038/nmeth.4267

Yang, X.; Liu, D.; Liu, F. et al (2013). HTQC: a fast quality control toolkit for Illumina sequencing data. BMC Bioinformatics 14, 33. Doi: 10.1186/1471-2105-14-33

Capítulo 2

CAPÍTULO 2

STR genetic diversity from the Human Genome Diversity Project (HGDP) populations.

Tamara Soledad Frontanilla*, Guilherme Valle-Silva, Jesús Ayala, Celso Teixeira Mendes-Junior.

[a]Departamento de Genética, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, 14049-900, Ribeirão Preto, SP, Brazil.

[b] Departamento de Química, Laboratório de Pesquisas Forenses e Genômicas, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, 14040-901, Ribeirão Preto, SP, Brazil.

[c] Universidad de la Integración de las Americas. Asunción, Paraguay.

*Corresponding author:
Tamara Soledad Frontanilla
tfronta@gmail.com

ORCID ID
Tamara Soledad Frontanilla: 0000-0002-6873-7813
Guilherme Valle-Silva: 0000-0002-0062-9162
Jesus Ayala: 0000-0002-7065-6879
Celso Teixeira Mendes-Junior: 0000-0002-7337-1203

ABSTRACT

The Human Genome Diversity Project (HGDP) is an international collaboration to create a database of diverse human populations. HGDP studied 54 worldwide populations comprising seven population groups: African, American, Central South Asian, East Asian, European, Middle Eastern, and Oceanian. This study aimed to perform a comprehensive genotyping analysis of STRs commonly used in forensic and population genetic studies from the Human Genome Diversity Project dataset and to publish it as an open-access STR database to contribute to future forensic genetics studies. A set of 22 STR markers were analyzed using high-coverage Whole-Genome Sequencing data from BAM files available at the International Genome Sample Resource. HipSTR was used to call genotypes from 929 samples from all 54 population samples. To validate our results, we directly compared our NGS-based and CE-based genotypes on 16 STRs available in Rosenberg lab (Stanford University) dataset. Also, the allele frequencies estimated were compared with the data stored at the SPSmart STR browser. Forensic parameters, allele frequencies, and Hardy-Weinberg equilibrium adherence were calculated for each population. Principal Coordinate Analysis (PCoA), the Analysis of Molecular Variance (AMOVA), and clustering analysis were used to evaluate population structure. The D21S11 marker could not be detected in the present study. The average successful calling rate was 90.27%, ranging from 58.56% (Penta D) to 97.85% (D3S1358). Comparing both databases, the average number of identical genotypes was 97.44%. All interpopulation genetic diversity analyses could differentiate the major biogeographic populations at both continental and subcontinental levels. In conclusion, this investigation offers a population genetics perspective based on a comprehensive genotyping analysis of STR commonly used in the forensic genetics field, concerning the whole Human Genome Diversity Project dataset. Except for Penta D and Penta E, all genotypes and allele frequencies presented in this study are supported by (a) previous reports that certify HipSTR's reliability, (b) the comparison between CE-derived and NGS-derived genotypes, (c) frequency data reports from worldwide populations, including the large pop.STR database, and (d) the conclusions achieved by our population genetics analysis that corroborates current knowledge regarding modern human demographic history.

Keywords: HipSTR; allele frequencies; forensic genetics; worldwide population; bioinformatics.

1. INTRODUCTION

The Human Genome Diversity Project (Almarri et al., 2020; Bergström et al., 2020; Cavalli-Sforza, 2005) (HGDP) is a collaboration of scientists worldwide to create a database of different world populations. It was started in 1990 by Stanford University's Morrison Institute (Cann et al., 2022; Rosenberg, 2006). The project initially had some ethical issues concerning indigenous populations (Dodson et al., 1999), who are considered vulnerable and might be exploited (Cavalli-sforza, 2005). In 1994, after a few years of discussion, the US National Research Council (NRC) of the National Academy of Sciences (NAS) recommended that the HGDP should proceed because of the countless scientific benefits, but always carrying out the necessary care and consent. This project studied 54 worldwide populations comprising seven population groups: African, American, Central South Asian, East Asian, European, Middle Eastern, and Oceanian (Bergström et al., 2020).

There are many international collaborative genome-wide studies, such as The Human Genome Project (HGP) (Birney, 2021) the Haplotype Map (HapMap) project (1000 Genomes Project Consortium et al., 2015), the Human Genome Diversity Project (Cavalli-Sforza, 2005), and the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015). The first two focus on mapping and sequencing genes to discover their relationship with different diseases. However, the HGDP and the 1000 Genomes Project are more interested in understanding the extent of genetic variation between humans. Although The 1000 Genomes Project has produced an extensive catalog of human genetic variation, the HGDP contains samples of underrepresented human populations or isolated indigenous populations that are necessary to better understand the demographic history and introgression of Neanderthals and Denisovans' DNA into modern human genomes (Callaway, 2019; Degioanni et al., 2019; Demeter et al., 2022).

Next generation sequencing (NGS) (Behjati et al., 2013), also known as massively parallel sequencing or deep sequencing, is a revolutionary technology that allows the sequencing of millions of small DNA fragments in parallel. Specifically developed bioinformatics tools are used to piece together these fragments using the human reference genome as a backbone. NGS can evaluate thousands or even millions of *loci* simultaneously compared with just a few dozen *loci* detected by PCR and electrophoresis (Bonneville *et al.*, 2020).

Genome-wide studies, including exome and/or whole-genome sequencing, are becoming more and more common worldwide for diagnosing rare genetic diseases and predicting possible forthcoming conditions. Such datasets allow for the analysis of more complex genetic regions that are usually left aside, such as Short Tandem Repeats (STR) markers. STRs are composed of consecutive repetitive units of 2-6 base pairs that form series with lengths of up to 100 nucleotides or even more (Fan; 2007). Typically, capillary electrophoresis (CE) is the technique used to genotype these markers after PCR amplification. However, recent articles (Ganschow et al., 2018; Gymrek et al., 2012; Valle-Silva et al., 2022; Warshauer et al., 2013; Willems et al., 2017) showed that specific bioinformatic tools could successfully genotype these markers from NGS data.

Haplotype inference and phasing for STRs (Willems et al., 2017; Gordon et al., 2017) (HipSTR) is a bioinformatic tool developed for calling STR markers specifically from Whole Genome Sequencing (WGS). It was created to process hundreds of samples at once, making it suitable to deal with large databases. Moreover, HipSTR learns locus-specific PCR stutter models using an EM algorithm, employing a specialized hidden Markov model to align reads to candidate alleles while accounting for STR artifacts and using phased SNP haplotypes to genotype and phase STR. HipSTR showed high accuracy in previous studies, demonstrating a 98.8% consistency compared with capillary electrophoresis in 118 samples (Halman; Oshlack, 2020). Valle-Silva et al. (2022) compared three software to genotype STR markers from NGS data showing more than 97% calling accuracy between them.

This study aimed to perform a comprehensive genotyping analysis of STRs commonly used in population genetics studies from the Human Genome Diversity Project dataset and to publish it as an open-access STR database.

2. METHODOLOGY

2.1 Genotype Calling

The population sample consisted of 929 individuals from the Human Genome Diversity Project (HGDP) panel, distributed across 54 worldwide populations that compose seven population groups: Africa (*n*=104), Americas (*n*=61), Central South Asia (*n*=197), East Asia (*n*=223), Europe (*n*=155), Middle East (*n*=161) and Oceania (*n*=28). These populations are described by Bergstrom et al. (Bergström et al., 2020) (Table 1). The CRAM files containing sequence data from these 929 samples are available at the International Genome Sample Resource, divided into two datasets: one presented by Mallick et al. (Mallick et al., 2016) (https://www.internationalgenome.org/data-portal/data-collection/hgdp), and the other by Bergström et al. (2020) (https://www.internationalgenome.org/data-portal/data-collection/sgdp).

Table 1. Population samples from the Human Genome Diversity Project (HGDP) used in this study (*n*=929).

Population	Subpopulation	Nomenclature	Number of individuals
Africa	BantuKenya	AFR001	11
	BantuSouthAfrica	AFR002	8
	Biaka	AFR003	22
	Mandenka	AFR004	22
	Mbuti	AFR005	13
	San	AFR006	6
	Yoruba	AFR007	22
America (Amerindians)	Colombian	AMR008	7
	Karitia	AMR009	12
	Maya	AMR010	21
	Pima	AMR011	13
	Surui	AMR012	8
Central/South Asia	Balochi	CSA013	24
	Brahui	CSA014	25
	Burusho	CSA015	24
	Hazara	CSA016	19
	Kalash	CSA017	22
	Makrani	CSA018	25
	Pathan	CSA019	24

	Sindhi	CSA020	24
	Uygur	CSA021	10
East Asia	0xi	EAS022	8
	Cambodian	EAS023	9
	Dai	EAS024	9
	Daur	EAS025	9
	Han	EAS026	33
	Hezhen	EAS027	9
	Japanese	EAS028	27
	Lahu	EAS029	8
	Miao	EAS030	10
	Mongolian	EAS031	9
	NorthernHan	EAS032	10
	Oroqen	EAS033	9
	She	EAS034	10
	Tu	EAS035	10
	Tujia	EAS036	9
	Xibo	EAS037	9
	Yakut	EAS038	25
	Yi	EAS039	10
Europe	Adygei	EUR040	16
	Basque	EUR041	23
	BergamoItalian	EUR042	12
	French	EUR043	28
	Orcadian	EUR044	15
	Russian	EUR045	25
	Sardinian	EUR046	28
	Tuscan	EUR047	8
Middle East	Bedouin	MES048	46
	Druze	MES049	42
	Mozabite	MES050	27
	Palestinian	MES051	46
Oceania	Bougainville	OCE052	11
	PapuanHighlands	OCE053	9
	PapuanSepik	OCE054	8

All samples were sequenced in high-coverage as described by Mallick et al. (MALLICK; LI; LIPSON; MATHIESON *et al.*, 2016) and Bergström et al. (2020). This coverage depth provides a reliable opportunity to genotype STR markers accurately despite their large sizes (i.e., repetitive sequences encompassing up to 130 bp).

A total of 22 *loci* commonly referenced in forensic practice were analyzed: CSF1PO, D1S1656, D2S441, D2S1338, D3S1358, D5S818, D7S820, D8S1179, D10S1248, D12S391, D13S317, D16S539, D18S51, D19S433, D21S11, D22S1045, FGA, Penta D, Penta E, TH01, TPOX, and vWA. The HipSTR (Willems et al., 2017) algorithm was run for each individual to genotype the 22 STRs based on the human reference genome GRCh38, and using the BED file hg38.hipstr_reference.bed with the flanking regions available in the HipSTR repository (https://github.com/HipSTR-Tool/HipSTR-references/). A minimum of 8 reads were accepted to obtain reliable genotypes, and the 15% stutter model as a calling filter was used.

The genotypes for each marker were calculated using three parameters provided in the output VCF file: the reference allele of each marker, the period (i.e., the length of each STR repeat unit), and the base pair differences (GB) in comparison with the reference allele (Willems et al., 2017). Nomenclature adjustments were made for D19S433, Penta D, Penta E, and vWA following Valle-Silva et al. (Valle-Silva et al., 2022) recommendations to couple with the nomenclature established by the International Society for Forensic Genetics (ISFG) (Gettings; et al., 2019). By using the IGV software (Robinson et al., 2017; Thorvaldsdóttir et al., 2013) and the HipSTR VizAln function (Willems et al., 2017), we have previously demonstrated that the alleles provided by HipSTR for these four markers (D19S433, Penta D, Penta E, and vWA) needed nomenclature adjustment to avoid a shift of some base pairs in allele calling (Valle-Silva et al., 2022). The adjustments consisted of removing two repeat units from all D19S433 and vWA alleles called by HipSTR, including one repeat unit into all Penta D alleles, and removing two nucleotides from all Penta E alleles (Gettings et al., 2019).

2.2 Statistical Analysis

Forensic parameters [Match Probability (MP), Power of Discrimination (PD), Power of Exclusion (PE), and Polymorphism Information Content (PIC)], allele frequencies, and Hardy-Weinberg equilibrium adherence were estimated for each population group using GenAlEx 6.5 (Peakall, 2012) and STRAF 2.5.1 software (Gouy; Zieger, 2017).

To explore the distribution of genetic diversity across populations of different ethnic backgrounds, the Principal Coordinate Analysis (PCoA), the Analysis of Molecular Variance (AMOVA), and clustering analysis were done using GenAlEx 6.5 (Peakall, 2012), Arlequin 3.5 (Excoffier; Lischer, 2010), and STRUCTURE 2.3.4 (Hubisz et al., 2009) software, respectively. The STRUCTURE analysis was performed for *k* ranging from 3 to 7, applying the correlated allele frequency model and 200.000 burn-in steps followed by 200.000 Markov Chain Monte Carlo interactions in 10 independent runs. The results from the runs with the largest "Estimated Ln Probability of Data" [LnP(D)] were selected and are depicted in bar plots created with Clumpak (Kopelman et al., 2015).

2.3 Genotype validation

We used two validation methodologies to verify the reliability of genotype data generated by HipSTR. The first one consisted of a direct comparison with CE-derived genotypes available in the Rosenberg's lab (Stanford University) dataset available at: https://rosenberglab.stanford.edu/data/algeehewittEtAl2016/HGDPmicrosatsIncludingCODIS.stru. The dataset is composed of the data published by Algee-Hewitt *et al.* (2016) (Algee-Hewitt et al., 2016) and Rosenberg et al. (2005) (Rosenberg et al., 2005). We used 865 individuals and 16 STR markers present in both the NGS and CE datasets for this validation.

In a secondary validation attempt, the allele frequencies estimated in the present study were compared with allele frequency data from the same seven major population groups (African, European, Middle Eastern, Central South Asian, East Asian, Oceanian, and American) stored at the SPSmart STR browser (Amigo et al., 2009; Fernandez et al., 2009) (Pop.STR). For this comparison, pairwise F_{ST} was estimated using the Arlequin software (Excoffier; Lischer, 2010).

3. RESULTS

STR genotypes established for each individual from the HGDP dataset using HipSTR are available in Supplementary Table 1 as an open-access database. The D21S11 marker was excluded because we failed in genotyping it. Figure 1 shows the genomic location of the D21S11 marker from an HGDP sample (HGDP01405) visualized with the IGV program. In this image we can see incomplete reads at the region, therefore, this could be the cause of HipSTR's failure to capture this marker. The mean coverage for genotype calling ranged from 29.765 (Penta D) to 53.869 (D3S1358) (Table 2).

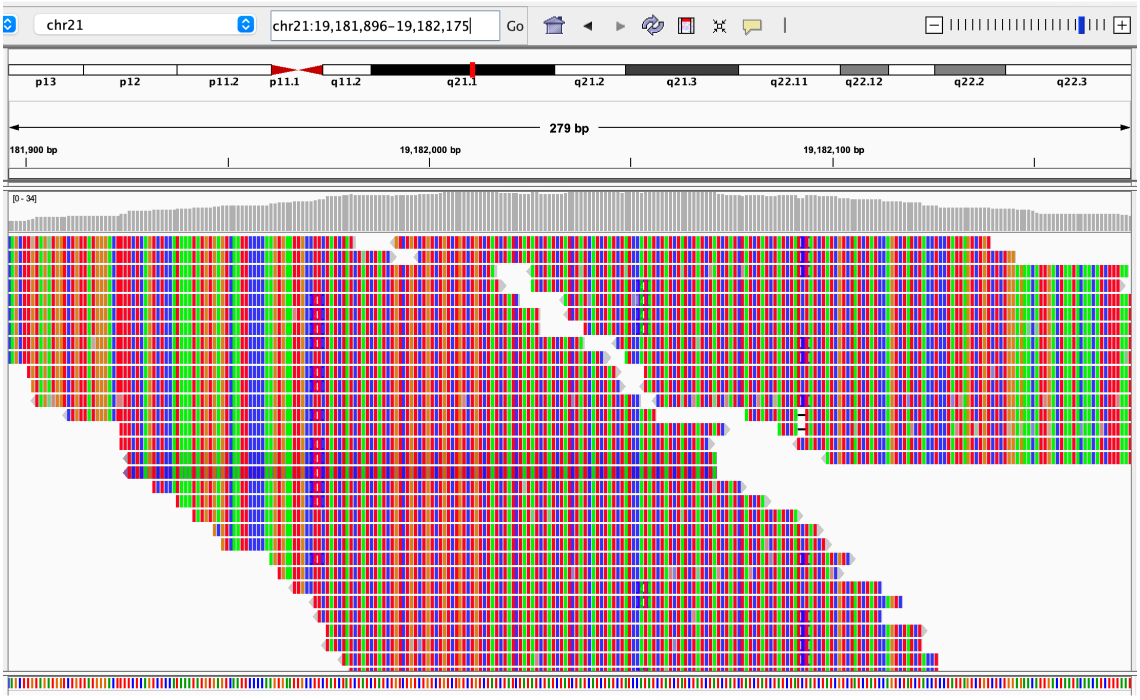

Figure 1. D21S11 genomic location of the HGDP01405 sample of the HGDP dataset.

Table 3 shows the allele frequencies and forensic parameters for the whole HGDP dataset, while Supplementary Table 2 presents these same parameters for each of the seven major population groups studied. The average successful calling rate was 90.27%, ranging from 58.56% (Penta D) to 97.85% (D3S1358) (Table 3). HipSTR failed in genotyping the Penta D alleles smaller than five repeats. Moreover, Penta E was in H-W disequilibrium in half (27) of the 54 population samples (Table 4). Thus, these markers were excluded from all interpopulation statistical analyses performed in the present study (Analysis of

Molecular Variance, STRUCTURE analysis, and PCoA). It is noteworthy that the D22S1045 was monomorphic in a small (*n*=13) Amerindian population sample of Mexico (Pima); however, this is due to a lack of genetic diversity in this locus rather than genotyping errors.

Table 2. Average coverages obtained for each STR using the HipSTR tool.

Maker	Lowest value	Median	Highest value	Mean	Standard deviation
CSF1PO	22	47	158	47.704	14.659
D1S1656	21	49	138	49.757	15.574
D2S441	19	48	125	48.643	14.656
D2S1338	28	52	115	47.994	22.810
D3S1358	23	53	134	53.869	15.041
D5S818	12	44	117	44.856	13.887
D7S820	18	41	118	41.309	12.603
D8S1179	24	50	137	50.962	15.589
D10S1248	20	43	105	43.073	14.102
D12S391	23	53	122	48.841	22.612
D13S317	15	40	120	40.670	13.406
D16S539	25	48	123	47.867	14.269
D18S51	29	52	154	50.016	22.480
D19S433	22	47	104	45.073	16.933
D22S1045	8	46	121	39.821	22.946
FGA	28	56	132	50.909	23.812
PentaD	16	38	130	29.765	27.295
PentaE	11	39	119	30.427	23.215
TH01	17	40	115	40.778	12.295
TPOX	19	37	101	35.364	14.620
vWA	23	44	173	43.086	22.804

Table 3. Allele frequencies and the forensic parameters estimated for each marker in the whole HGDP dataset.

Allele	CSF1PO	D1S1656	D2S441	D2S1338	D3S1358	D5S818	D7S820	D8S1179	D10S1248	D12S391	D13S317	D16S539	D18S51	D19S433	D22S1045	FGA	PentaD	PentaE	TH01	TPOX	vWA
5																	0.002	0.100			
5.2														0.001							
6	0.001						0.001										0.007		0.204	0.012	
7	0.010					0.019	0.018		0.001		0.001						0.014	0.108	0.237	0.006	
8	0.012	0.003	0.001			0.014	0.167	0.006	0.002		0.150	0.013					0.040	0.053	0.141	0.457	
9	0.024	0.002	0.003			0.055	0.086	0.005	0.001		0.099	0.172		0.001			0.252	0.036	0.265	0.135	
9.1			0.005				0.001														
9.3			0.001																0.133		
10	0.261	0.006	0.215			0.132	0.256	0.110	0.001		0.087	0.116	0.005	0.014	0.014		0.165	0.085	0.018	0.076	
10.1			0.001																		
10.3									0.001												
11	0.275	0.071	0.351			0.314	0.277	0.063	0.008		0.269	0.291	0.014	0.012	0.179		0.194	0.232	0.002	0.276	0.001
11.1							0.001														
11.2													0.001	0.001							
11.3			0.056																		
11.4																		0.001			
12	0.356	0.085	0.087		0.001	0.288	0.162	0.117	0.052		0.284	0.262	0.087	0.072	0.018		0.132	0.205	0.001	0.038	
12.1												0.001									
12.2													0.001	0.008							
12.3			0.009											0.001							
13	0.054	0.095	0.032		0.003	0.163	0.030	0.235	0.258		0.084	0.125	0.128	0.259	0.005		0.134	0.088			0.001
13.1													0.001								
13.2													0.002	0.044							
13.3		0.001	0.002																		
14	0.006	0.117	0.211		0.057	0.014	0.003	0.238	0.296	0.001	0.026	0.018	0.181	0.286	0.046		0.040	0.057			0.131
14.2														0.060							
14.3		0.002																			
15	0.001	0.183	0.023		0.351	0.001		0.166	0.223	0.022	0.001	0.002	0.157	0.093	0.326		0.009	0.029			0.087

15.2										0.089					
15.3	0.021														
16	0.185	0.003	0.027	0.284	0.001	0.048	0.124	0.023	0.121	0.035	0.251	0.001	0.007	0.006	0.230
16.2				0.001					0.001	0.023					
16.3	0.043														
17	0.064	0.001	0.138	0.201		0.011	0.033	0.092	0.126	0.001	0.154	0.001	0.002		0.268
17.1								0.001							
17.2										0.001					
17.3	0.080							0.008							
18	0.011		0.081	0.094		0.002	0.002	0.210	0.089		0.007	0.012			0.188
18.2										0.001		0.001			
18.3	0.023							0.016							
19			0.163	0.009				0.205	0.051		0.001	0.053			0.078
19.1								0.001							
19.2								0.001				0.002			
19.3	0.008							0.006							
20			0.136					0.141	0.021			0.078			0.014
20.2												0.001			
21			0.036					0.099	0.010			0.122			0.001
21.2												0.003			
22			0.053					0.081	0.004			0.219			
22.2												0.005			
22.3								0.001							
23			0.226					0.056	0.001			0.158			
23.2												0.004			
24			0.086					0.027	0.001			0.178			
24.2								0.001				0.003			
24.3												0.001			
25			0.045					0.007				0.102			
25.2												0.004			
26			0.009					0.001				0.041			

26.2										0.001											
27																0.008					
28																0.004					
28.1																0.001					
30																0.001					
	CSF1PO	**D1S1656**	**D2S441**	**D2S1338**	**D3S1358**	**D5S818**	**D7S820**	**D8S1179**	**D10S1248**	**D12S391**	**D13S317**	**D16S539**	**D18S51**	**D19S433**	**D22S1045**	**FGA**	**PentaD**	**PentaE**	**TH01**	**TPOX**	**vWA**
N	905	903	908	793	909	907	907	903	886	806	899	902	823	852	778	800	544	627	905	850	803
Na	10	18	16	11	9	10	11	11	13	22	9	9	20	19	10	24	13	12	8	7	10
Ho	0.728	0.843	0.675	0.755	0.724	0.731	0.763	0.792	0.721	0.814	0.750	0.763	0.854	0.764	0.666	0.754	0.778	0.440	0.706	0.664	0.762
He	0.726	0.883	0.773	0.864	0.744	0.770	0.795	0.829	0.777	0.864	0.800	0.787	0.877	0.822	0.773	0.860	0.832	0.859	0.794	0.689	0.808
MP	0.126	0.025	0.081	0.037	0.109	0.085	0.071	0.051	0.081	0.033	0.068	0.076	0.028	0.053	0.087	0.036	0.050	0.058	0.070	0.151	0.062
PE	0.473	0.681	0.391	0.519	0.466	0.478	0.532	0.584	0.462	0.625	0.509	0.532	0.703	0.534	0.377	0.516	0.558	0.140	0.438	0.374	0.531
PD	0.874	0.975	0.919	0.963	0.891	0.915	0.929	0.949	0.919	0.967	0.932	0.924	0.972	0.947	0.913	0.964	0.950	0.942	0.930	0.849	0.938
PIC	0.677	0.873	0.741	0.850	0.702	0.735	0.765	0.807	0.742	0.850	0.772	0.755	0.864	0.801	0.738	0.844	0.811	0.845	0.762	0.642	0.781
CMP	**3.72.E-26**																				
CPE	0.999999676																				

N: number of samples; Na: number of alleles; Ho: Observed Heterozygosity; He: Expected Heterozygosity; MP: match probability; PE: power of exclusion; PD: power of discrimination; PIC: polymorphism information content; CMP: combined match probability; CPE combined power of exclusion.

Table 4. Probabilities of adherence to Hardy-Weinberg equilibrium proportions for each STR in all 54 subpopulations analyzed in the HGDP. Significant *p*-values (α = 0.05) are in boldface.

Pop	CSF1PO	D1S1656	D2S441	D2S1338	D3S1358	D5S818	D7S820	D8S1179	D10S1248	D12S391	D13S317	D16S539	D18S51	D19S433	D22S1045	FGA	PentaD	PentaE	TH01	TPOX	vWA
AFR001	0.539	0.248	0.766	0.628	0.673	0.433	0.763	0.988	0.290	0.174	0.837	0.300	0.521	0.682	0.340	0.223	0.423	0.336	0.191	0.505	0.842
AFR002	0.712	0.350	0.824	0.229	0.824	0.930	0.440	0.621	0.374	0.888	0.273	0.261	0.438	0.161	**0.017**	0.177	0.157	0.116	0.673	0.704	0.390
AFR003	0.074	0.812	**0.008**	0.636	0.181	0.537	0.529	0.900	0.947	0.379	0.808	0.929	0.562	0.363	**0.042**	0.787	0.450	0.059	0.553	0.834	0.985
AFR004	0.742	0.461	0.662	0.201	0.916	0.967	0.576	0.630	0.964	0.441	0.724	0.648	0.141	0.992	0.245	0.864	0.526	**0.004**	0.702	0.136	0.603
AFR005	0.207	0.256	0.085	**0.012**	0.617	0.938	0.519	0.427	0.669	0.751	0.631	0.518	0.289	0.353	0.256	0.121	0.823	**0.038**	0.569	0.607	0.900
AFR006	0.556	0.548	0.227	0.654	0.868	0.678	0.556	0.868	0.166	0.874	0.054	0.226	0.174	0.626	0.767	0.062	0.766	0.207	0.995	0.502	0.393
AFR007	0.291	0.317	0.981	0.305	0.059	**0.011**	0.381	0.289	0.605	0.385	0.425	0.257	0.812	0.631	0.708	0.283	0.595	**0.007**	0.682	0.149	0.561
AMR008	0.914	0.678	0.914	0.694	0.466	0.950	0.312	0.556	**0.021**	0.735	0.479	0.776	0.511	0.575	**0.034**	0.282	0.387	0.116	0.626	0.152	0.626
AMR009	0.122	0.724	0.197	0.308	0.785	0.942	0.950	0.711	0.615	0.535	**0.044**	0.568	**0.031**	0.763	0.451	0.820	0.841	0.083	0.376	0.792	**0.043**
AMR010	0.999	0.446	0.990	0.687	0.524	0.254	0.707	0.403	0.461	**0.001**	0.283	0.795	0.149	0.210	**0.011**	0.138	0.678	**0.008**	0.795	0.463	0.566
AMR011	0.800	0.645	0.199	0.461	0.637	**0.001**	0.229	0.853	0.337	0.949	0.307	0.615	0.460	0.200	-	0.546	0.711	**0.003**	0.623	**0.028**	0.827
AMR012	0.820	0.498	0.686	0.978	0.983	**0.028**	0.719	0.836	0.726	0.557	0.947	0.217	0.545	0.117	0.054	0.628	0.542	**0.046**	0.733	0.409	0.442
CSA013	0.532	0.273	0.165	0.350	**0.000**	0.215	0.869	0.718	0.478	0.306	0.915	0.224	0.737	0.063	0.565	**0.011**	0.892	**0.000**	0.655	0.861	0.802
CSA014	0.909	0.407	0.073	0.334	**0.010**	0.809	0.973	0.688	0.920	0.870	0.588	0.735	0.620	0.106	0.557	0.329	0.363	0.614	0.325	0.726	0.845
CSA015	0.615	0.611	0.617	0.746	**0.043**	**0.037**	0.695	0.144	0.875	0.135	0.922	0.374	0.999	0.091	0.765	**0.000**	0.584	**0.011**	0.366	0.797	0.342
CSA016	0.652	0.940	**0.009**	0.521	0.849	0.669	0.180	0.917	**0.006**	0.457	0.290	0.073	0.764	0.867	0.163	**0.039**	0.272	0.136	0.194	0.423	0.611
CSA017	**0.873**	0.966	**0.079**	0.799	0.467	0.985	0.486	0.983	0.608	0.840	0.632	0.361	0.577	0.064	0.001	**0.088**	0.607	0.022	0.882	0.823	0.810
CSA018	0.759	0.565	0.997	0.920	0.442	0.847	0.876	0.645	0.320	**0.030**	0.154	0.249	0.733	0.192	0.221	0.596	0.248	**0.001**	0.912	0.949	0.343
CSA019	0.984	0.787	0.908	0.488	0.593	0.716	0.085	0.857	0.465	**0.007**	0.350	0.144	0.510	0.716	0.845	0.345	0.559	0.153	0.764	0.834	0.908
CSA020	0.976	0.585	0.797	0.792	0.436	0.707	0.124	0.689	0.939	0.930	0.795	0.168	0.191	0.107	0.937	0.485	0.081	**0.001**	0.955	0.461	0.684
CSA021	0.093	0.803	0.884	0.585	0.379	0.777	0.264	0.899	0.606	0.609	0.756	0.258	0.214	0.678	0.678	0.084	0.509	**0.013**	**0.029**	0.540	0.399
EAS022	0.460	0.611	0.421	0.262	0.587	0.440	0.760	0.269	0.896	0.718	0.154	0.741	0.453	0.466	0.570	0.505	0.396	**0.019**	0.311	0.572	0.212
EAS023	0.779	0.917	0.173	0.732	0.231	0.543	0.676	0.493	0.656	0.779	0.511	0.523	0.469	0.902	0.516	0.245	0.744	0.109	0.103	0.895	0.521
EAS024	0.213	0.607	0.182	**0.006**	0.948	0.134	0.544	0.101	0.947	0.866	0.685	0.467	0.233	0.172	0.320	0.899	0.849	**0.003**	0.837	0.896	0.744
EAS025	0.083	0.621	0.423	0.656	0.878	0.197	**0.016**	**0.009**	0.620	0.326	0.137	0.276	0.482	0.312	0.320	0.388	0.338	**0.006**	0.586	0.322	0.815
EAS026	0.930	0.823	0.831	0.868	0.960	0.693	0.569	0.727	0.121	0.239	0.920	**0.025**	0.989	0.961	0.610	0.540	0.562	**0.000**	0.587	0.867	0.829
EAS027	0.677	0.103	0.969	0.373	0.652	0.794	0.479	0.577	0.183	0.413	0.206	0.524	0.932	0.661	0.720	0.704	0.804	0.229	0.099	0.726	0.518

EAS028	0.712	**0.008**	0.215	0.815	0.481	0.792	0.243	0.257	0.690	0.988	0.674	0.649	0.301	0.404	0.850	0.928	0.373	**0.000**	0.223	0.623	0.622
EAS029	0.631	0.572	0.423	0.651	0.947	0.777	0.850	0.820	0.478	0.725	0.389	0.147	0.739	0.353	0.796	0.649	0.641	0.132	0.824	0.685	0.226
EAS030	0.374	0.839	0.345	0.354	**0.019**	0.567	0.714	0.452	0.317	0.432	0.695	0.832	0.247	0.922	0.538	0.469	0.540	0.256	0.329	0.606	0.738
EAS031	0.376	0.967	0.799	0.641	0.638	0.878	0.922	0.223	0.895	0.529	0.734	0.575	0.168	0.676	0.572	0.375	0.891	0.062	0.941	0.789	0.761
EAS032	0.202	0.581	0.571	0.192	0.398	0.704	0.288	0.198	0.717	0.418	0.548	0.595	0.154	0.332	0.815	0.956	0.870	0.103	0.497	0.111	0.402
EAS033	0.791	0.389	0.268	0.626	0.599	0.387	0.895	0.789	0.615	0.602	0.639	0.877	0.252	0.120	0.581	0.459	0.936	**0.022**	0.806	0.946	0.465
EAS034	0.546	0.287	0.363	0.440	0.506	0.481	0.375	0.787	0.974	0.966	0.800	0.925	0.582	0.538	0.274	0.258	0.507	0.177	0.374	0.120	0.570
EAS035	0.681	0.413	0.590	0.637	0.379	0.427	0.537	0.924	0.611	0.497	0.453	**0.050**	0.816	0.431	0.799	0.459	0.711	0.227	0.501	0.587	0.534
EAS036	0.119	0.171	0.369	0.402	0.656	0.685	0.544	0.956	0.855	0.804	0.402	0.922	0.854	0.232	0.366	0.451	0.677	0.125	0.723	0.472	0.370
EAS037	0.265	0.276	0.342	0.412	0.532	0.509	0.187	0.077	0.812	0.442	0.752	0.382	0.279	0.768	0.544	0.078	0.716	0.282	0.552	0.716	0.891
EAS038	0.977	**0.005**	0.801	0.400	0.928	0.902	0.924	0.354	0.951	0.942	0.871	0.592	0.966	0.964	0.238	0.791	0.646	0.215	0.768	0.849	0.866
EAS039	0.769	0.354	0.564	**0.003**	0.393	0.617	0.798	0.617	0.581	0.147	0.883	0.273	0.879	0.163	0.154	0.615	0.134	0.062	0.064	0.856	0.780
EUR040	0.641	0.462	0.385	0.918	0.369	0.487	0.564	0.235	0.804	0.796	0.983	0.953	0.168	0.917	**0.016**	0.488	0.658	**0.002**	0.430	0.084	0.793
EUR041	0.440	0.326	0.469	0.743	0.563	**0.033**	0.727	0.722	0.841	0.705	0.253	0.800	0.472	0.810	0.988	0.140	0.593	0.118	0.606	0.941	0.735
EUR042	0.091	0.073	0.292	0.376	0.434	0.470	**0.005**	0.218	0.851	0.525	**0.026**	0.512	0.689	0.540	0.907	0.221	0.220	**0.032**	0.216	0.625	0.245
EUR043	0.730	0.709	0.153	0.341	0.305	0.970	0.100	0.809	0.920	**0.049**	0.139	0.653	0.833	**0.001**	0.089	0.839	0.214	0.053	0.936	0.655	0.671
EUR044	**0.005**	0.396	0.607	0.413	0.093	0.258	**0.045**	0.772	0.833	0.666	0.080	0.966	0.170	0.950	0.966	0.721	0.402	**0.038**	0.345	0.425	0.086
EUR045	0.774	0.312	0.732	**0.018**	0.889	0.432	0.747	0.304	0.335	0.551	0.209	0.087	0.689	0.504	**0.042**	0.693	0.841	**0.000**	0.503	0.998	0.241
EUR046	0.515	0.137	0.065	0.728	0.530	0.393	0.665	0.223	0.414	0.104	0.822	0.961	0.421	0.692	0.680	0.331	0.172	**0.002**	0.619	0.333	0.234
EUR047	0.792	0.496	0.307	0.227	0.977	0.340	0.834	0.326	0.878	0.569	0.735	0.502	0.550	0.104	0.436	0.177	0.371	0.086	0.848	0.949	0.851
MES048	0.145	0.890	0.061	0.068	0.907	**0.008**	0.364	0.840	0.511	**0.021**	0.440	0.635	0.256	0.238	0.753	**0.009**	0.944	**0.000**	0.781	0.519	0.641
MES049	0.144	**0.000**	1.000	0.270	0.230	0.864	0.689	**0.000**	0.484	**0.000**	0.618	0.976	0.226	0.374	**0.006**	0.232	0.377	**0.000**	0.115	0.085	0.282
MES050	0.342	0.566	0.241	0.851	0.667	0.230	0.478	0.123	0.865	0.462	0.990	0.648	0.081	0.948	0.714	0.499	0.556	**0.021**	0.728	**0.050**	0.685
MES051	0.973	**0.024**	0.276	0.628	0.950	0.921	**0.005**	0.357	0.954	0.987	0.746	0.076	0.570	0.760	0.658	**0.000**	0.599	**0.023**	0.814	0.974	0.833
OCE052	0.636	0.867	0.181	0.432	0.463	0.857	0.979	0.214	0.472	0.780	0.420	0.263	0.594	0.566	0.635	0.164	0.565	0.085	0.678	0.377	0.175
OCE053	0.254	0.451	0.719	0.111	0.799	0.671	0.864	0.633	0.936	0.338	0.552	0.517	0.964	0.764	0.244	0.361	0.558	**0.000**	0.381	0.557	0.453
OCE054	0.849	0.154	0.686	0.674	0.974	0.062	0.272	0.276	0.647	0.160	0.183	0.412	0.371	0.916	0.108	0.164	0.729	0.157	0.757	0.677	0.867

The Principal Coordinates Analysis (PCoA) shows a good differentiation between major biogeographic populations at both continental (Figure 2) and subcontinental (Figure 3) scales. For Figure 2, all subpopulations were grouped, revealing four different population clusters. As expected, the African, Amerindian, and Oceanian populations were placed separately (in different quadrants), while the European and Asian populations were clustered together, revealing a similar genetic composition. The two principal coordinates account for 70.42% of the variance. In Figure 3, although the two first coordinates account for only 24.14% of the variance, the distribution of the 54 subpopulations was consistent with what was observed in Figure 2, resulting in four different and well-defined clusters. However, in the cluster with the European and Asian populations, one may observe an overlapping of populations from the four groups, mainly European, Middle Eastern, and Central South Asian populations, corroborating their shared ancestry and similar genetic compositions.

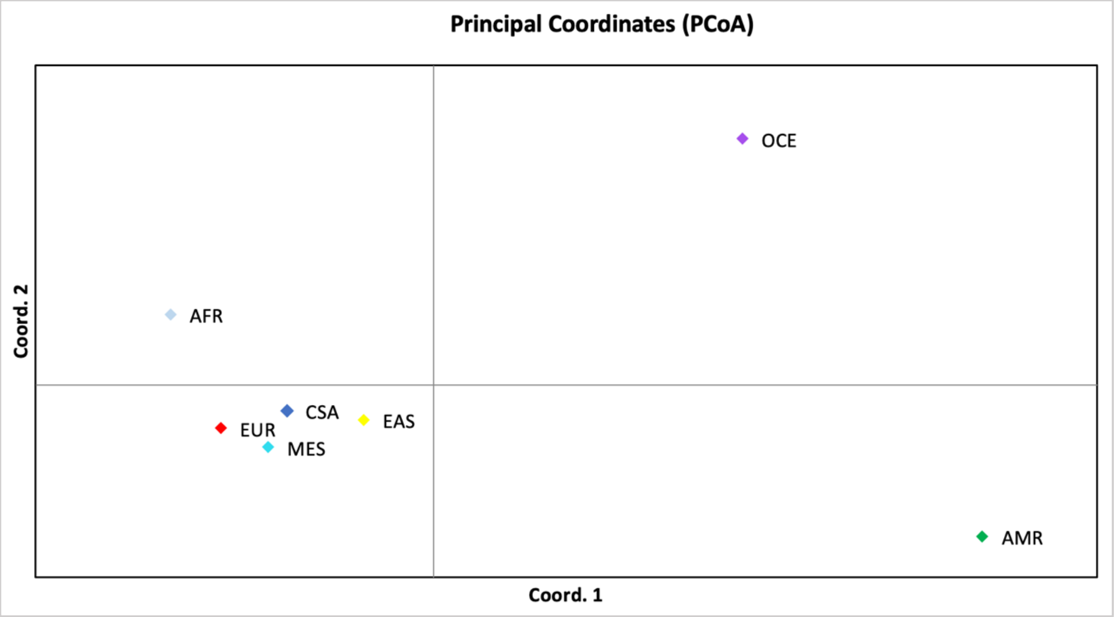

Figure 2. Principal Coordinates Analysis (PCoA) based on autosomal STR data from the 7 major populations of the HGDP. Coordinates 1 and 2 account for 40.66% and 29.76% of the variance, respectively. Penta D and Penta E markers were excluded from this analysis. (AFR: African; CSA: Central South Asia; EAS: East Asia; EUR: European; MES: Middle East; OCE: Oceania).

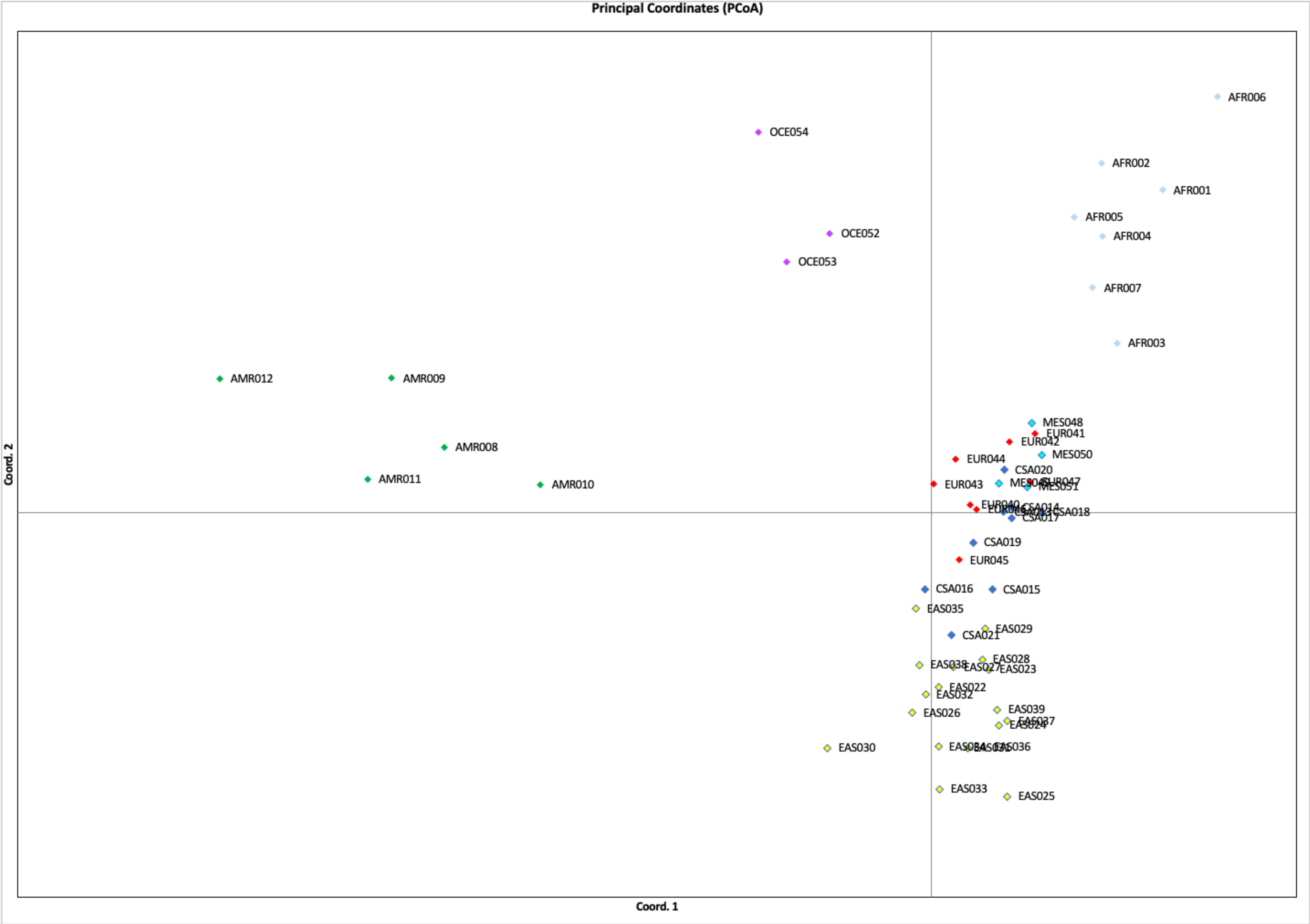

Figure 3. Principal Coordinates Analysis (PCoA) based on autosomal STR data from the 54 sub-populations of the HDGP. Coordinates 1 and 2 account for 13.48% and 10.66% of the variance, respectively. Penta D and Penta E markers were excluded from this analysis. (AFR: African; CSA: Central South Asia; EAS: East Asia; EUR: European; MES: Middle East; OCE: Oceania).

Similar results were obtained with the STRUCTURE analysis. Figure 4 depicts STRUCTURE results from runs obtained with *k* ranging from 3 to 7. With *k* = 5, one may observe that African, Amerindian and Oceanian groups mainly present their own clusters, while European and Asian populations display their shared ancestry, especially European, Middle Eastern and Central South Asian populations. By analyzing *k* = 7, it is possible to observe that the Central South Asian populations are highly heterogeneous with each other but also present evident differences when compared to European and Middle Eastern populations. Although minor differences arise with *k* = 7, European and Middle Eastern populations are very similar in all *k*.

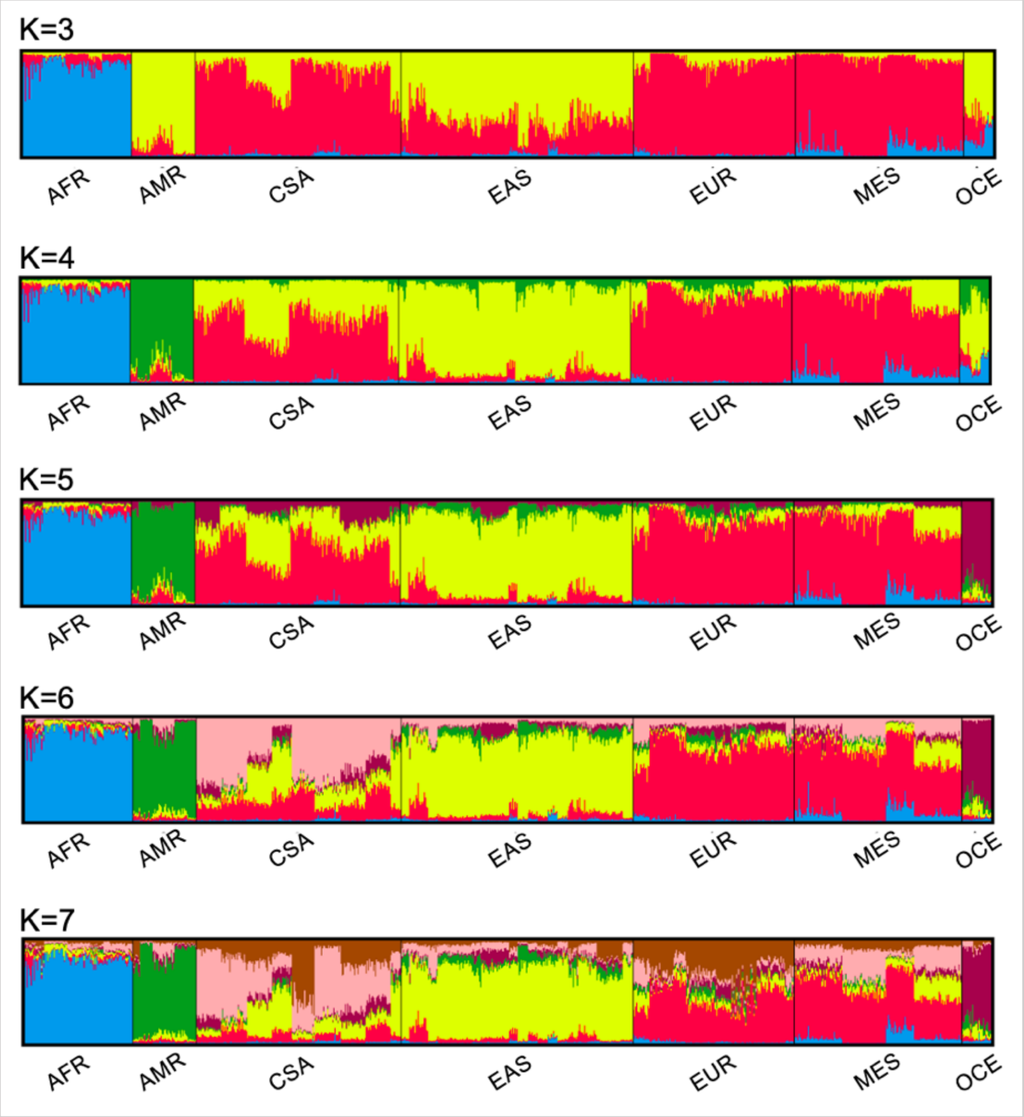

Figure 4. STRUCTURE analysis based on autosomal STR data from the 54 subpopulations of HGDP. Seven sets of 10 independent runs with the number

of clusters ranging from 3 to 7 were conducted. Each bar plot depicts the results from the run with the largest LnP(D) for the given *k*. Penta D and Penta E markers were excluded from this analysis. (AFR: African; CSA: Central South Asia; EAS: East Asia; EUR: European; MES: Middle East; OCE: Oceania).

To verify the distribution of variance in different levels, an AMOVA was performed assuming a hierarchical structure gathering the populations in seven groups: AFR, AMR, CSA, EAS, EUR, MES and OCE without the Penta E and Penta D markers. Most of the variance is observed within populations (95.84%). Differences between the seven groups account for 2.61% of the variance, whereas only 1.55% of the variance occurs due to differences between populations from the same group. An alternative structure, composed of only four groups (merging CSA, EAS, EUR and MES populations in a single group), revealed an increase in the variance between groups: differences between the four groups account for 4.14% of the variance, whereas only 2,17% of the variance occurs due to differences between populations from the same group and as expected most of the variance is observed within populations (93.67%).

The genotypes calculated with HipSTR were compared with a dataset of previously obtained CE-derived genotypes (Algee-Hewitt et al., 2016; Rosenberg et al., 2005). The average number of identical genotypes was 97.44% (median = 99.35%) (Supplementary Table 3), ranging from 88.25% (FGA) to 99.88% (D8S1179). Given the high proportion of genotypes with only one correct allele for some *loci*, these figures are much better when the assignment of correct alleles are taken into account: the average number of correct alleles was 98.49% (median = 99.67%), ranging from 92.12% (FGA) to 99.94% (D8S1179) (Supplementary Table 3). The errors in allele assignment are summarized in Supplementary Table 4. Inconsistencies were considered as “stutter-related” errors when HipSTR failed to detect the smaller allele in situations in which the CE-derived genotype indicated a heterozygote composed of contiguous alleles (e.g., 11/12) and called it as a false homozygous (e.g., 12/12). Stutter-related errors accounted for 13.2% of all errors. Other types of inconsistencies were considered as “stutter-unrelated” errors (86.8%). All 311 errors are detailed in Supplementary Table 5.

Allele frequencies estimated from the HGDP dataset were also compared with the frequencies presented in the SPSmart STR browser (Amigo *et al.*, 2009)

(Pop.STR) using F_{ST} (Table 5). For this comparison, we used the same subpopulations that are present in both databases. The Penta E marker presented *p*-values lower than 0.05 in all groups except in the Middle East (MES). In general, we observed only 10 significant F_{ST} spread out in four markers: D2S441, D5S818, Penta D, and Penta E.

Table 5. Probabilities obtained by F_{ST} analysis of population differentiation based on genotype frequencies of each STR, comparing population groups from the Human Genome Diversity Project with those from the SPSmart STR browser (Pop.STR). Significant *p*-values ($\alpha = 0.05$) are in boldface. The probabilities that remain significant after the Bonferroni correction for multiple tests ($\alpha_{BONFERRONI} = 0.05/147 = 0.00034$) are also underlined.

Marker	AFR		AMR		CSA		EAS		EUR		MES		OCE	
	F_{ST}	*p*-value	F_{ST}	*p*-value	F_{ST}	*p*-value	F_{ST}	*p*-value	F_{ST}	*p*-value	F_{ST}	*p*-value	F_{ST}	*p*-value
CSF1PO	-0.00915	0.99980+-0.0001	-0.01250	0.85556+-0.0036	-0.00428	0.96010+-0.0021	-0.00433	0.99921+-0.0003	-0.00627	0.99584+-0.0006	-0.00614	0.99871+-0.0004	-0.02731	0.90872+-0.0032
D1S1656	-0.00938	0.99999+-0.0000	-0.01514	0.99994+-0.0000	-0.00492	0.99999+-0.0000	-0.00427	0.99998+-0.0000	-0.00616	0.99999+-0.0000	-0.00593	0.99999+-0.0000	-0.03287	0.99999+-0.0000
D2S441	-0.00905	0.99999+-0.0000	-0.01479	0.97297+_0.0125	-0.00472	0.99792+-0.0001	-0.00412	0.99736+-0.0002	0.00523	0.11839+-0.0010	0.03174	**0.00063+-0.0001**	-0.03320	0.94017+-0.0008
D2S1338	0.00212	0.27614+-0.0014	-0.01342	0.97511+-0.0005	-0.00061	0.51765+-0.0017	0.00286	0.12064+-0.0010	0.00151	0.26167+-0.0014	-0.00443	0.96563+-0.0006	-0.03561	0.99999+-0.0000
D3S1358	-0.00958	0.99999+-0.0000	-0.01559	0.99697+-0.0002	-0.00480	0.99841+-0.0001	-0.00442	0.99999+-0.0000	-0.00627	0.99912+-0.0001	-0.00605	0.99853+-0.0001	-0.03566	0.99999+-0.0000
D5S818	-0.00902	0.99961+-0.0001	-0.01539	0.99631+-0.0002	-0.00496	0.99965+-0.0001	-0.00428	0.99908+-0.0001	-0.00644	0.99999+-0.0000	0.04888	**0.00000+-0.0000**	-0.03525	0.99999+-0.0000
D7S820	-0.00842	0.99382+-0.0002	-0.01511	0.99693+-0.0002	-0.00487	0.99962+-0.0001	-0.00443	0.99999+-0.0000	-0.00646	0.99999+-0.0000	-0.00570	0.99429+-0.0002	-0.02634	0.92461+-0.0008
D8S1179	-0.00953	0.99999+-0.0000	-0.01601	0.99999+-0.0000	-0.00500	0.99999+-0.0000	-0.00440	0.99999+-0.0000	-0.00632	0.99996+-0.0000	-0.00603	0.99995+-0.0000	-0.03535	0.99999+-0.0000
D10S1248	-0.00920	0.99997+-0.0000	-0.01472	0.99562+-0.0002	-0.00393	0.94223+-0.0007	-0.00385	0.97001+-0.0005	-0.00599	0.99106+-0.0003	-0.00569	0.99402+-0.0002	-0.03438	0.98741+-0.0003
D12S391	-0.00846	0.99995+-0.0000	-0.01439	0.99089+-0.0003	-0.00448	0.99985+-0.0000	-0.00247	0.84891+-0.0011	-0.00443	0.98000+-0.0004	-0.00459	0.98615+-0.0003	-0.03569	0.99999+-0.0000
D13S317	-0.00879	0.99329+-0.0003	-0.01328	0.96635+-0.0006	-0.00487	0.99981+-0.0000	-0.00432	0.99956+-0.0001	-0.00626	0.99988+-0.0000	-0.00605	0.99948+-0.0001	-0.03247	0.99041+-0.0003
D16S539	-0.00948	0.99992+-0.0000	-0.01540	0.99538+-0.0002	-0.00498	0.99999+-0.0000	-0.00418	0.99545+-0.0002	-0.00613	0.99891+-0.0001	-0.00616	0.99991+-0.0000	-0.03353	0.99790+-0.0001
D18S51	-0.00885	0.99994+-0.0000	-0.01508	0.99976+-0.0000	-0.00367	0.97798+-0.0005	-0.00281	0.91754+-0.0009	-0.00611	0.99997+-0.0000	-0.00585	0.99999+-0.0000	-0.01898	0.87383+-0.0010
D19S433	-0.00866	0.99983+-0.0000	-0.01513	0.99990+-0.0000	-0.00453	0.99955+-0.0001	-0.00324	0.93432+-0.0008	-0.00596	0.99906+-0.0001	-0.00596	0.99992+-0.0000	-0.03597	0.99999+-0.0000
D22S1045	0.00985	0.07334+-0.0008	-0.00716	0.55926+-0.0015	-0.00301	0.79675+-0.0013	-0.00370	0.95798+-0.0006	-0.00576	0.98184+-0.0004	-0.00513	0.97348+-0.0005	0.05184	0.06208+-0.0008
FGA	0.00366	0.19120+-0.0013	-0.01013	0.93542+-0.0008	-0.00156	0.67142+-0.0014	-0.00038	0.47693+-0.0016	-0.00438	0.94444+-0.0007	-0.00525	0.99428+-0.0003	-0.02768	0.99177+-0.0003
Penta D	0.02162	**0.00154+-0.0001**	-0.01285	0.94995+-0.0008	-0.00167	0.63971+-0.0015	-0.00248	0.83218+-0.0012	-0.00383	0.86665+-0.0011	-0.00469	0.97591+-0.0005	-0.02859	0.98105+-0.0004
Penta E	0.00953	**0.03937+-0.0006**	0.02947	**0.00501+-0.0002**	0.01702	**0.00010+-0.0000**	0.05375	**0.00000+-0.0000**	0.00603	**0.04632+-0.0007**	0.00332	0.13552+-0.0011	0.04362	**0.01032+-0.0003**
TH01	-0.00952	0.99999+-0.0000	-0.01577	0.99999+-0.0000	-0.00477	0.99722+-0.0002	-0.00437	0.99976+-0.0000	-0.00587	0.98292+-0.0004	-0.00551	0.98326+-0.0004	-0.03620	0.99999+-0.0000
TPOX	-0.00954	0.99987+-0.0000	-0.01540	0.99478+-0.0002	-0.00495	0.99929+-0.0001	-0.00432	0.99626+-0.0002	-0.00514	0.91091+-0.0009	-0.00617	0.99977+-0.0000	-0.03418	0.98061+-0.0004
vWA	-0.00788	0.98409+-0.0004	-0.01544	0.99814+-0.0001	-0.00498	0.99998+-0.0000	-0.00420	0.99766+-0.0001	-0.00590	0.99440+-0.0002	-0.00595	0.99843+-0.0001	-0.03169	0.99310+-0.0003

4. DISCUSSION

This study offers a STR database from high-coverage next-generation sequencing data derived from the 54 population samples that compose the Human Genome Diversity Project (HGDP).

Accurate STR genotyping from NGS data has been challenging due to the high sequencing error rates and difficulties in aligning repetitive sequences (Fungtammasan et al., 2015). However, Bornman et al. (2012) demonstrated that CODIS *loci* could be accurately called even from complex mixtures using an NGS approach. Notwithstanding that, capillary electrophoresis (CE) is, until now, and will continue to be for a long time, the most used technique to genotype STRs due to its simplicity. CE doesn't offer nucleotide sequence information (Bornman et al., 2012), while an NGS assay allows differentiating isometric alleles (isoalleles), which would permit to increase forensic informativeness (i.e., power of discrimination and power of exclusion) (Hert et al., 2008). In this study, HipSTR was used to differentiate alleles by size, and not by sequence due to the large number of samples processed simultaneously.

HipSTR presented some problems in specific markers like D21S11, Penta D, and Penta E. The D21S11 marker was excluded because HipSTR couldn't genotype it. The same problem has already been reported by Valle-Silva et al. (Valle-Silva et al., 2022) in a previous study. D21S11 is a complex marker, which can cause alignment errors (Rockenbauer et al., 2014). In figure one, we demonstrate incomplete reads in the genomic location of the D21S11 marker, and this could be the reason why HipSTR failed to capture this marker. Since HipSTR managed to capture the D21S11 marker in a previous amplicon-based study, the problem could be related to the alignment and not to the bioinformatics tool used (Valle-Silva; et al., 2022). Moreover, the length of the sequenced alleles may also play a part, given that even the smallest common D21S11 allele (with 26 repeats encompassing 104 nucleotides) is large, and sequencing error rates increase with STR length (Kelkar et al., 2008). These issues may lead to mapping failure during the alignment step.

On the other hand, we may have failed in genotyping small Penta D alleles that presented less than 5 repeats. This situation mainly affected the African populations:

many studies, including the pop.STR data (Amigo et al., 2009), show that Penta D has a very high frequency of the 2.2 allele (0.20%). Also, in this study Penta D presented the lowest successful calling rate (58.56%). Penta E deviated from H-W equilibrium in 27 of 54 populations, being responsible for more than 30% of Hardy-Weinberg departures observed. Because of these problems, Penta D and Penta E were excluded from all interpopulation statistical analyses (Analysis of Molecular Variance, PCoA and clustering analysis). We don't recommend using these markers for population genetics or human identification purposes using the HipSTR software. However, toaSTR (Carsten, 2017; Ganschow et al., 2018; Valle-Silva; et al., 2022) showed very effective Penta D and Penta E genotyping in previous studies. The limitation of this software, which prevented its use in the present study, is that it can only process one sample at a time, while HipSTR can process thousands of samples in parallel.

The D22S1045 marker showed to be monomorphic in an Amerindian population from México, Pima. This population is considered to be composed of descendants of the ancient Hohokam, who have inhabited the Sonoran Desert and Sierra Madre regions for centuries. Today, they are present in two countries, in the USA (Arizona state), as "*The O'odham*", and in Mexico as "*O'ob"* or *"Pima Bajo*" (Schulz et al., 2015). According to the most recent data from the Mexican government, currently, 1.540 Pima exist in the country (HOPE, 2006). The SPSmart STR browser (Amigo et al., 2009) (Pop.STR) revealed precisely the same situation, with only allele 15 being observed. The Pop.STR studied 14 individuals from this population, while HGDP sampled 13 individuals. Small populations typically show a high rate of inbreeding, which produces the fixation of some alleles (Hartl, 2020).

When the genotypes calculated with HipSTR were compared with those from the dataset provided by Algee-Hewitt et al. (2016) and Rosenberg et al. (2005), the average number of identical genotypes was 97.44% (median = 99.35%). The FGA and D22S1045 STRs were the most problematic ones and strongly influenced the average. In the case of FGA, one of the reasons may be the length of some alleles. For instance, although alleles with more than 30 repeats are extremely rare, the largest described FGA allele is composed of 51 tetranucleotide repeats. The observed stutter-unrelated problems could be related to the positioning of flanking regions, tri-allelic patterns or alignment errors. Although it is not reasonable to assign all the inconsistencies to problems in the NGS-based procedure, particularly given that in the two CE-based studies mentioned above were in fact large-scale genome-wide studies that prevented a careful evaluation of each genotype for 1160 STRs in 2034 subjects from worldwide populations, it is noteworthy that HipSTR uses previously obtained bam files. Thus, additional efforts in improving the WGS alignment procedure, particularly considering the repetitive nature of microsatellite regions, may increase the overall accuracy of genotype calling by the HipSTR algorithm. Unfortunately, CE-derived genotypes were unavailable for five markers (D1S1656, D2S1338, D12S391, Penta D, and Penta E), rendering the secondary validation attempt involving the comparison of allele frequencies using pairwise F_{ST} of utmost importance to assess the reliability of their NGS-based genotypes.

The Principal Coordinates Analysis (PCoA) was able to separate the major populations correctly (Figure 2) and also, the sub-populations (Figure 3). Similar results were revealed by the clustering analysis (Figure 4). While African, Amerindian and Oceanian populations are clearly differentiated, Asian (CSA, EAS, MES) and European (EUR) populations present high levels of shared ancestry. Although modern humans arose in Africa, the Middle East is considered the cradle of Eurasian civilization (Guest; Sahebkar, 2021), where the world's first civilizations originated. Thanks to its economic supremacy, Europe ended up colonizing the Middle East and leaving a large immigrant community. This situation could be the reason for the genetic similarity between the individuals from these regions. Historically, Central Asia has been an intersection between Western and Eastern Eurasian people, leading to the current high levels of genetic admixture and diversity (González-Ruiz et al., 2012).

The Structure analysis (Figure 4) shows that the African and Native American populations form largely distinct homogeneous clusters, while the Middle Eastern, European, Central, and South Asian populations form a more heterogeneous cluster. These findings reflect the more isolated nature of the former populations and corroborate the idea that although forensic STRs do show relatively low F_{ST}, their high heterozygosities strengthens their capacity to uncover patterns of population clustering, also revealed by other sets of markers (JOBLING, 2022). Our findings agree with the data presented by Pemberton et al. regarding human microsatellite variation on large databases, including the HGDP-CEPH (Pemberton et al., 2013).

Despite the problems already discussed, HipSTR proved to be highly effective for genotyping STR markers from NGS data, mainly for CODIS markers which are the most used in the forensic area. Notwithstanding, we recommend using more than one software to genotype these markers from NGS to obtain high efficiency and circumvent the genotype calling issues we have described.

5. CONCLUSION

In conclusion, this investigation offers a population genetics perspective based on a comprehensive genotyping analysis of standard STR used in the forensic genetics field concerning the whole Human Genome Diversity Project. Penta D and Penta D Markers were excluded from our analysis because they did not show up as reliable markers. All the remaining genotypes and allele frequencies presented in this study are supported by (a) previous reports that certify HipSTR's reliability, (b) the comparison between CE-derived and NGS-derived genotypes, (c) frequency data reports from worldwide populations, including the large pop.STR database, and (d) the conclusions achieved by our population genetics analysis that corroborates current knowledge regarding modern human demographic history.

6. FUNDING

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. C.T.M.J. (#312802/2018-8) is supported by a Research Fellowship from CNPq/Brazil.

7. CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest.

8. COMPLIANCE WITH ETHICAL STANDARDS

Not applicable.

9. REFERENCES

Algee-Hewitt, B. F.; Edge, M. D.; Kim, J.; Li, J. Z. et al (2016). Individual Identifiability Predicts Population Identifiability in Forensic Microsatellite Markers. Curr Biol 26(7):935-942. Doi: 10.1016/j.cub.2016.01.065

Almarri MA, Bergström A, Prado-Martinez J, Yang F., et al (2020). Population Structure, Stratification, and Introgression of Human Structural Variation. Cell. 9;182(1):189-199. Doi: 10.1016/j.cell.2020.05.024.

Amigo J.; Christopher, P.; Toño, S.; Fernandez, F.L., et al (2009). pop.STR—An online population frequency browser for established and new forensic STRs. Forensic Sci. Int. Genet. Suppl. Ser. 2, 361–362. Doi: 10.1016/j.fsigss.2009.08.178

Behjati S, Tarpey PS (2013). What is next generation sequencing? Arch Dis Child Educ Pract 98(6):236-8. Doi: 10.1136/archdischild-2013-304340.

Bergström A, McCarthy SA, Hui R, Almarri MA., et al (2020). Insights into human genetic variation and population history from 929 diverse genomes. Science 20;367(6484). Doi: 10.1126/science.aay5012

Birney E (2021). The International Human Genome Project. Hum Mol Genet. 1,30(R2):R161-R163. Doi: 10.1093/hmg/ddab198.

Bonneville R, Krook MA, Chen HZ, Smith A., et al (2020). Detection of Microsatellite Instability Biomarkers via Next-Generation Sequencing. Methods Mol Biol. 2055:119-132. Doi: 10.1007/978-1-4939-9773-2_5.

Bornman, D.M, Hester, M.E., Schuetter, J.M.; Kasoji, M.D., et al (2012). Short-read, high-throughput sequencing technology for STR genotyping. Biotech. Rapid Dispatches 1–6. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4301848/

Callaway E (2019). First portrait of mysterious Denisovans drawn from DNA. Nature 573(7775):475-476. Doi: 10.1038/d41586-019-02820-0.

Cann HM, de Toma C, Cazes L, Legrand MF., et al (2002) A human genome diversity cell line panel. Science 12;296(5566):261-2. Doi: 10.1126/science.296.5566.261b.

Cavalli-Sforza LL (2005). The Human Genome Diversity Project: past, present and future. Nat Rev Genet. 6(4):333-40. Doi: 10.1038/nrg1596.

Degioanni A, Bonenfant C, Cabut S, Condemi S (2019). Living on the edge: Was demographic weakness the cause of Neanderthal demise? PLoS One 29;14(5):e0216742. Doi: 10.1371/journal.pone.0216742.

Demeter, F.; Zanolli, C.; Westaway, K. E.; Joannes-Boyau, R. et al (2022). A Middle Pleistocene Denisovan molar from the Annamite Chain of northern Laos. Nat Commun, 13(1):2557.

Dodson M, Williamson R (1999). Indigenous peoples and the morality of the Human Genome Diversity Project. J Med Ethics 25(2):204-8. Doi: 10.1136/jme.25.2.204.

Excoffier, L.; Lischer, H.E (2010). Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Mol. Ecol. Resour. 10, 564–567. Doi: 10.1111/j.1755-0998.2010.02847.x

Fan H, Chu JY (2007). A brief review of short tandem repeat mutation. Genomics Proteomics Bioinformatics 5(1):7-14. Doi: 10.1016/S1672-0229(07)60009-6.

Fungtammasan, A.; Ananda, G.; Hile, S.E.; Su, M.S., et al (2015). Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. Genome Res. 25, 736–749. Doi: 10.1101/gr.185892.114.

Ganschow, S.; Silvery, J.; Kalinowski, J.; Tiemann, C (2018). toaSTR: A web application for forensic STR genotyping by massively parallel sequencing. Forensic Sci. Int. Genet. 37, 21–28. Doi: 10.1016/j.fsigen.2018.07.006.

Gettings, K.B.; Ballard, D.; Bodner, M.; Borsuk, L.A.., et al (2019). Report from the STRAND Working Group on the 2019 STR sequence nomenclature meeting. Forensic Sci. Int. Genet. 43, 102165. Doi: 10.1016/j.fsigen.2019.102165.

González-Ruiz M, Santos C, Jordana X, Simón M., et al (2012). Tracing the origin of the east-west population admixture in the Altai region (Central Asia). PLoS One 7(11):e48904. Doi: 10.1371/journal.pone.0048904.

Gouy, A.; Zieger, M (2017). STRAF-A convenient online tool for STR data evaluation in forensic genetics. Forensic Sci. Int. Genet. 30, 148–151. Doi: 10.1016/j.fsigen.2017.07.007.

Guest PC, Sahebkar A (2021). Research in the Middle East into the Health Benefits of Curcumin. Adv Exp Med Biol.1291:1-13. Doi: 10.1007/978-3-030-56153-6_1.

Gymrek, M.; Golan, D.; Rosset, S.; Erlich, Y (2012). lobSTR: A short tandem repeat profiler for personal genomes. Genome Res. 22, 1154–1162. Doi: 10.1101/gr.135780.111.

Halman, A.; Oshlack, A (2020). Accuracy of short tandem repeats genotyping tools in whole exome sequencing data. F1000Res 9, 200. Doi: 10.12688/f1000research.22639.1

Hartl, D (2020). A Primer of Population Genetics and Genomics. 4a ed. Oxford University Press.

Hert DG, Fredlake CP, Barron AE (2008). Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. Electrophoresis 29(23):4618-26. Doi: 10.1002/elps.200800456.

Hope, M (2006). Pueblos Indígenas del México Contemporáneo. 2006. Available from: https://www.inpi.gob.mx/2021/dmdocuments/pimas.pdf.

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009). Inferring weak population structure with the assistance of sample group information. Mol Ecol Resour 9(5):1322-32. Doi: 10.1111/j.1755-0998.2009.02591.x.

Jobling, M.A (2022). Forensic genetics through the lens of Lewontin: Population structure, ancestry and race. Philos. Trans. R. Soc. Lond. B Biol. Sci. 2022, 377, 20200422. Doi: 10.1098/rstb.2020.0422

Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD (2008). The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Res 18(1):30-8. Doi: 10.1101/gr.7113408.

Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA., et al (2015). Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Mol Ecol Resour 15(5):1179-91. doi: 10.1111/1755-0998.12387.

Mallick S, Li H, Lipson M, Mathieson I., et al (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 13;538(7624):201-206. Doi: 10.1038/nature18964.

Peakall, R.; Smouse, P.E (2012). GenAlEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research-an update. Bioinformatics 28, 2537–2539. Doi: 10.1093/bioinformatics/bts460

Pemberton TJ, DeGiorgio M, Rosenberg NA (2013). Population structure in a comprehensive genomic data set on human microsatellite variation. G3 Bethesda 20;3(5):891-907. Doi: 10.1534/g3.113.005728

Robinson, J.T.; Thorvaldsdóttir, H.; Wenger, A.M.; Zehir, A., et al (2017). Variant Review with the Integrative Genomics Viewer. Cancer Res. 77, e31–e34. Doi: 10.1158/0008-5472.CAN-17-0337

Rockenbauer E, Hansen S, Mikkelsen M, Børsting C., et al (2014). Characterization of mutations and sequence variants in the D21S11 locus by next generation sequencing. Forensic Sci Int Genet 8(1):68-72. Doi: 10.1016/j.fsigen.2013.06.011.

Rosenberg NA (2006). Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet. 70(Pt 6):841-7. Doi: 10.1111/j.1469-1809.2006.00285.x.

Rosenberg NA, Mahajan S, Ramachandran S, Zhao C., et al (2005). Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet 1(6):e70. Doi: 10.1371/journal.pgen.0010070.

Schulz LO, Chaudhari LS (2015). High-Risk Populations: The Pimas of Arizona and Mexico. Curr Obes Rep. 4(1):92-8. Doi: 10.1007/s13679-014-0132-9

Thorvaldsdóttir, H.; Robinson, J.T.; Mesirov, J.P (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief. Bioinform. 14, 178–192. Doi: 10.1093/bib/bbs017.

Valle-Silva, G.; Frontanilla, T.S.; Ayala, J.; Donadi, E.A., et al (2022). Analysis and comparison of the STR genotypes called with HipSTR, STRait Razor and toaSTR by using next generation sequencing data in a Brazilian population sample. Forensic Sci. Int. Genet. 58, 102676. Doi: 10.1016/j.fsigen.2022.102676

Warshauer, D.H.; Lin, D.; Hari, K.; Jain, R., et al (2013). STRait Razor: A length-based forensic STR allele-calling tool for use with second generation sequencing data. Forensic Sci. Int. Genet. 7, 409–417. Doi: 10.1016/j.fsigen.2013.04.005

Willems, T.; Zielinski, D.; Yuan, J.; Gordon, A., et al (2017). Genome-wide profiling of heritable and de novo STR variations. Nat. Methods 14, 590–592. Doi: 10.1038/nmeth.4267

CAPÍTULO 3

CAPÍTULO 3

Open-Access Worldwide Population STR Database Constructed Using High-Coverage Massively Parallel Sequencing Data Obtained from the 1000 Genomes Project

Tamara Soledad Frontanilla[1*], Guilherme Valle-Silva[2], Jesus Ayala[3], Celso Teixeira Mendes-Junior[2].

1. Departamento de Genética, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto 14049-900, SP, Brazil
2. Departamento de Química, Laboratório de Pesquisas Forenses e Genômicas, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto 14040-901, SP, Brazil
3. Facultad de Ingeniería Informática, Universidad de la Integración de las Americas, Asunción 00120-6, Paraguay

* These authors contributed equally to this work.

Doi: https://doi.org/10.3390/genes13122205

Supplementary Materials:

https://www.mdpi.com/article/10.3390/genes13122205/s1

Received: 15 October 2022

Revised: 13 November 2022

Accepted: 21 November 2022

Published: 24 November 2022

ABSTRACT

Achieving accurate STR genotyping by using next-generation sequencing data has been challenging. To provide the forensic genetics community with a reliable open-access STR database, we conducted a comprehensive genotyping analysis of a set of STRs of broad forensic interest obtained from 1000 Genome populations. We analyzed 22 STR markers using files of the high-coverage dataset of Phase 3 of the 1000 Genomes Project. We used HipSTR to call genotypes from 2504 samples obtained from 26 populations. We were not able to detect the D21S11 marker. The Hardy-Weinberg equilibrium analysis coupled with a comprehensive analysis of allele frequencies revealed that HipSTR was not able to identify longer alleles, which resulted in heterozygote deficiency. Nevertheless, AMOVA, a clustering analysis that uses STRUCTURE, and a Principal Coordinates Analysis showed a clear-cut separation between the four major ancestries sampled by the 1000 Genomes Consortium. Except for larger Penta D and Penta E alleles, and two very small Penta D alleles (2.2 and 3.2) usually observed in African populations, our analyses revealed that allele frequencies and genotypes offered as an open-access database are consistent and reliable.

Keywords: HipSTR; allele frequencies; forensic genetics; worldwide population; bioinformatics.

1. INTRODUCTION

Next-generation sequencing (NGS), also known as massively parallel or deep sequencing, is a technology that allows millions of DNA fragments to be sequenced in parallel. NGS can deal with several regions or targets simultaneously, enabling variation sites or mutations in the genome to be detected. This technology has allowed worldwide human genetic diversity to be studied for various purposes, including forensic human identification (Børsting et al., 2015; Alvarez-Cubero et al., 2017; Ballard et al., 2020).

Advances in the genomics area have made it possible to use NGS techniques in a more accessible way, mostly because of lower costs. Currently, many researchers are performing whole-exome (WES) and even whole-genome (WGS) sequencing to estimate polygenic risk scores and probabilities of developing multifactorial diseases associated with various genetic regions at once, which would be a more laborious and costly issue if using traditional methodologies (Børsting et al., 2015).

The 1000 Genomes Consortium is a worldwide collaboration that has produced an extensive catalog of human genetic variation. The consortium has sequenced whole genomes of 2504 individuals belonging to multiple populations derived from five population groups: African, East Asian, European, South Asian, and admixed Americans (1000 Genomes Project Consortium et al., 2015). These data are freely available at the International Genome Sample Resource website (https://www.internationalgenome.org) to generate a variant call format file that uses a set of specific command lines (Clarke et al., 2017). In 2015, during Phase 3 of the Project, the consortium analyzed the genomes of all the individuals by using a combination of low-coverage whole-genome sequencing (WGS), deep exome sequencing, and dense microarray genotyping. The consortium described worldwide patterns of genomic diversity on the basis of Single Nucleotide Polymorphisms (SNPs), indels, and structural variants (SVs), including deletions, insertions, duplications, inversions, and copy-number variants (CNVs), but it did not analyze or study short tandem repeat (STR) markers in depth (Sudmant et al., 2015).

STR markers are crucial in human identification. These markers have high polymorphism levels and are particularly useful for interpreting mixtures of biological samples. However, in addition to the issue of small-sized amplicons, genotyping STR markers by using NGS data is difficult because alignment and stutter errors are frequent (Fungtammasan et al., 2015). Achieving accurate genotyping by employing NGS data has been challenging because these data have high sequencing error rates (Bornman et al., 2012). Gymrek et al. (2012) managed to obtain and to analyze STR markers from the dataset of the 1000 Genomes Project using lobSTR (Gymrek et al., 2012). Given that high coverage is mandatory for reliable STR genotype calling to be achieved, a primary concern regarding that study was that the data obtained from the 1000 Genomes Project available for lobSTR were generated by employing shallow sequencing coverage (2x–6x), so the calling was potentially susceptible to errors (Willems et al., 2017).

To circumvent this coverage issue, the New York Genome Center (NYGC) recently re-sequenced the 2504 samples of the panel of Phase 3 of the 1000 Genomes Project with high (30x) coverage, and aligned the sequence data to GRCh38. These publicly available data could be used to call STR markers reliably (Clarke et al., 2017; Fairley et al., 2020).

NGS technology allows dozens of STR markers to be analyzed together with different classes of markers that provide complementary contributions to population genetics and human identification. For example, including SNPs used as predictors of ancestry and phenotypic characteristics into commercial kits that employ capillary electrophoresis is unfeasible, but they can be combined with STR markers in NGS assays (Børsting et al., 2015). The problem with the NGS technology is the large amount of data generated and the lack of bioinformatic tools to analyze it (Børsting et al., 2015). Some tools (e.g., lobSTR (Gymrek et al., 2012), STRait Razor (Warshauer et al., 2013), toaSTR (Ganschow et al., 2008), and HipSTR (Willems et al., 2017), among others) were developed to analyze STR markers by using NGS data. Each tool employs different algorithms and flanking regions to capture STR reads.

Haplotype inference and phasing for STRs (HipSTR) was developed for calling microsatellites specifically from WGS Illumina FASTq files. HipSTR was designed to deal with genotyping errors and to obtain more robust STR genotypes. HipSTR accomplished this by learning locus-specific PCR stutter models, with the aid of an EM algorithm, by employing a specialized hidden Markov model to align reads to candidate alleles while accounting for STR artifacts, and by using phased SNP haplotypes to genotype and to phase STR markers. These factors turned HipSTR into one of the most reliable tools for genotyping STRs from Illumina sequencing data (Willems et al, 2017; Valle-Silva et al., 2022).

In contrast to other tools, HipSTR can process hundreds of samples at once. It also allows the user to determine the set of STR markers that must be analyzed and the flanking regions that must be used to capture them. In fact, previous studies showed that HipSTR provides accurate genotype calling. HipSTR accuracy was tested by comparing WGS calls from 118 samples to capillary electrophoresis data, which resulted in 98.8% consistency (Willems et al, 2017; Halman et al., 2020). Recently, we compared HipSTR with Strait Razor and toaSTR, to find that the three tools present high allele calling accuracy (greater than 97%) (Valle-Silva et al., 2022). Although data processing with HipSTR is more complex and requires bioinformatics knowledge and some nomenclature adjustments, this tool is currently the fastest and most appropriate to deal with larger datasets, including whole genomes (Valle-Silva et al., 2022).

In this investigation we conducted a comprehensive genotyping analysis of a set of STRs of broad forensic interest obtained from the 1000 Genomes populations, aiming to release a reliable open-access STR database that should contribute to future studies in the field of forensic genetics.

2. MATERIALS AND METHODS

2.1. Genotype Calling

Genotypes were called from 2504 individuals belonging to 26 populations derived from five population groups analyzed by the 1000 Genomes Consortium, namely African (AFR), East Asian (EAS), European (EUR), South Asian (SAS), and admixed American (AMR) (1000 Genomes Project Consortium et al., 2015). The NYGC re-sequenced the samples of Phase 3 of the 1000 Genomes Project in a high-coverage (30x) assay by applying the NovaSeq 6000 Sequencing System (Illumina, Inc.; San Diego, CA, USA) with a paired-end approach (2 × 150 bp). Then, the NYGC made the data freely available at: https://www.internationalgenome.org/data-portal/datacollection/30x-grch38.

We used CRAM files to obtain the STR genotypes with the aid of the HipSTR software (Willems et al., 2017). We selected 22 autosomal microsatellites that are commonly used in forensic practice: CSF1PO, D1S1656, D2S441, D2S1338, D3S1358, D5S818, D7S820, D8S1179, D10S1248, D12S391, D13S317, D16S539, D18S51, D19S433, D21S11, D22S1045, FGA, Penta D, Penta E, TH01, TPOX, and vWA.

To genotype the 22 STR markers based on the human reference genome GRCh38, we ran the HipSTR algorithm for each individual. For this purpose, we used a BED file with the coordinates of each STR region of interest, which was available in the HipSTR repository (Willems et al., 2017) (https://hipstr-tool.github.io/HipSTR-tutorial/; accessed on 10 July 2021) as described elsewhere (Valle-Silva et al., 2022). We applied the calling filter (15% stutter model) and a minimum of eight reads to obtain more reliable genotypes. According to a binomial distribution, this minimum number of reads ensures ($p > 0.99$) that a homozygous genotype is called because of lack of variability at a given locus and not because the second allele has not been sampled.

To perform genotype calling, we used the VCF output file produced by HipSTR and took three parameters into account: the reference allele of each marker, the period (i.e., the length of each STR repeat unit), and the base pair differences (GB) as compared to the reference allele. We adjusted the nomenclature for D19S433, Penta D, Penta E, and vWA by following the recommendations made by Valle-Silva et al. (Valle-Silva et al., 2022): removal of two repeat units from all D19S433 and vWA alleles called by HipSTR, inclusion of one repeat unit into all Penta D alleles, and removal of two nucleotides from all Penta E alleles. By using IGV software 2.8.2 (Thorvaldsdóttir et al., 2013; Robinson et al., 2022) and the HipSTR VizAln function (Willems et al., 2017), Valle-Silva et al. (2022) demonstrated that such adjustments are necessary to prevent some base pairs from shifting in allele calling when compared to the nomenclature established by the ISFG (Gettings et al., 2019).

2.2. Statistical Analysis

We calculated allele frequencies, the Hardy-Weinberg equilibrium, and forensic parameters [Match Probability (MP), Power of Discrimination (PD), Power of Exclusion (PE), and Polymorphism Information Content (PIC)] for each population sample or each population group using GenAlEx 6.5 (Peakall et al., 2012) and STRAF 2.5.1 (Gouy et al., 2017) software.

We employed Principal Coordinates Analysis (PCoA) using GenAlEx (Peakall et al., 2012), Analysis of Molecular Variance (AMOVA) using Arlequin (Excoffier et al., 2010), and clustering analyses using STRUCTURE 2.3.4 (Hubisz et al., 2009) to explore how genetic diversity is distributed across populations of different ethnic backgrounds. We performed STRUCTURE analysis for k ranging from 3 to 6 by applying the correlated allele frequencies model, 100,000 burn-in steps followed by 100,000 Markov Chain Monte Carlo interactions, in 100 independent runs. We selected the results from the runs with the largest "Estimated Ln Probability of Data" [LnP (D)] and depicted them in bar plots created with Distruct 1.1 (Rosenberg, 2004).

We also compared the allele frequencies estimated from the 1000 Genomes Project dataset with STR data retrieved from the same five major population groups (African, European, East Asian, South Asian, and admixed American) that compose the SPSmart STR browser (PopSTR) (Amigo et al., 2009). For this purpose, we employed Arlequin software to compare the allele frequencies of each STR marker for a given population group between the two datasets by using F_{ST} and an exact test of population differentiation based on genotype frequencies (Excoffier et al., 2010). We made this comparison to verify the reliability of genotype data generated by HipSTR.

3. RESULTS

The STR genotypes defined for each individual from the newest dataset released by the 1000 Genomes Project are available in Supplementary Table S1 as an open-access database. We excluded the D21S11 marker because we did not succeed in genotyping it (See discussion). Apart from this marker, the mean coverage for calling genotypes ranged from 37.14 (TPOX) to 52.53 (D12S391) (Table 1). The average successful calling rate was 98.59%; this rate ranged from 84.18% (Penta E) to 100% (CSF1PO, D2S441, D2S1338, D3S1358, D5S818, D8S1179, D22S1045, and TPOX) (Table 2).

Table 1. Average coverages obtained for each STR using the HipSTR tool.

Marker	Lowest value	Median	Highest value	Mean	Standard deviation
CSF1PO	21	44	91	44.54	8.29
D1S1656	24	45	92	45.48	8.58
D2S441	26	49	131	49.17	9.04
D2S1338	28	50	105	51.28	9.64
D3S1358	28	51	119	51.73	9.33
D5S818	20	42	98	42.90	8.14
D7S820	20	38	86	39.00	7.78
D8S1179	26	47	96	48.09	8.95
D10S1248	18	40	100	40.75	7.99
D12S391	26	52	113	52.53	9.47
D13S317	11	37	79	37.93	7.50
D16S539	21	44	92	44.70	8.49
D18S51	24	47	91	47.38	9.07
D19S433	19	45	89	45.28	8.62
D22S1045	22	49	111	49.47	9.11
FGA	23	50	118	51.33	9.49
Penta D	19	43	95	43.71	8.74
Penta E	18	41	107	41.49	8.04
TH01	16	40	83	40.73	8.01
TPOX	15	37	86	37.14	7.66
vWA	21	47	105	48.44	9.36

Table 2. Allelic frequencies and the forensic parameters estimated for each marker in the whole 1000 Genomes dataset.

Allele	CSF1PO	D1S1656	D2S441	D2S1338	D3S1358	D5S818	D7S820	D8S1179	D10S1248	D12S391	D13S317	D16S539	D18S51	D19S433	D22S1045	FGA	Penta D	Penta E	TH01	TPOX	vWA
5											0.0002						0.0125	0.0861	0.0020		
6							0.0004		0.0002					0.0004			0.0034	0.0017	0.1905	0.0210	
7	0.0198	0.0002				0.0118	0.0178		0.0002		0.0018	0.0004		0.0002			0.0150	0.1084	0.2828	0.0074	
7.3	0.0002																				
8	0.0218	0.0080	0.0006	0.0002		0.0222	0.1842	0.0064	0.0006		0.1394	0.0314		0.0002	0.0006		0.0539	0.0854	0.1287	0.4217	
8.3																			0.0002		
9	0.0334	0.0004	0.0018		0.0004	0.0421	0.0942	0.0066	0.0004		0.0891	0.1805	0.0004	0.0012	0.0002		0.2195	0.0358	0.2314	0.1534	
9.2			0.0052				0.0016													0.0002	
9.3																			0.1491		
10	0.2398	0.0080	0.2264			0.0976	0.2534	0.0953	0.0010		0.0763	0.1123	0.0046	0.0028	0.0154		0.1660	0.0873	0.0150	0.0621	0.0004
10.2			0.0004				0.0004						0.0006								
10.3	0.0002																		0.0002		
11	0.2658	0.0788	0.3401		0.0004	0.3187	0.2506	0.0753	0.0166		0.2650	0.2888	0.0110	0.0210	0.1649		0.1839	0.1774	0.0002	0.2971	0.0014
11.2							0.0004				0.0002			0.0008							
11.3			0.0529																		
11.4																	0.0002				
12	0.3375	0.0732	0.1062		0.0022	0.3097	0.1650	0.1180	0.0684		0.2912	0.2336	0.0805	0.0745	0.0170		0.1506	0.1886		0.0359	0.0008
12.2							0.0002						0.0006	0.0122							
12.3			0.0022											0.0004							
13	0.0693	0.1142	0.0282	0.0002	0.0034	0.1841	0.0275	0.2340	0.2616		0.0982	0.1317	0.1155	0.2706	0.0034		0.1380	0.1065		0.0006	0.0072
13.2													0.0032	0.0341							
13.3			0.0008						0.0002				0.0002								
14	0.0108	0.1472	0.2121	0.0006	0.0777	0.0116	0.0042	0.2418	0.2816	0.0010	0.0370	0.0200	0.1674	0.2726	0.0509		0.0396	0.0669		0.0006	0.1255
14.2													0.0010	0.0623			0.0002				0.0004
14.3		0.0028	0.0004										0.0002								
15	0.0016	0.1772	0.0202	0.0014	0.3075	0.0020		0.1566	0.2220	0.0336	0.0016	0.0014	0.1670	0.1042	0.3313	0.0004	0.0134	0.0380			0.1217
15.1														0.0002							
15.2					0.0006								0.0004	0.0799							0.0002
15.3		0.0272							0.0002												
16		0.1410	0.0026	0.0318	0.3031	0.0002		0.0559	0.1178	0.0342			0.1414	0.0295	0.2502	0.0006	0.0031	0.0176			0.2251
16.2					0.0002					0.0004				0.0240		0.0004					
16.3		0.0560											0.0002								
17		0.0448		0.1328	0.2063			0.0080	0.0268	0.1109			0.1187	0.0046	0.1472	0.0016	0.0007	0.0005			0.2433
17.2		0.0004								0.0014				0.0034							
17.3		0.0796								0.0078											
18		0.0060		0.0897	0.0903			0.0022	0.0022	0.2264			0.0787		0.0170	0.0110					0.1753
18.2										0.0008			0.0002	0.0008		0.0030					
18.3		0.0298								0.0096						0.0002					0.0002
19		0.0006		0.1679	0.0072				0.0002	0.1743			0.0519		0.0014	0.0673					0.0770
19.2										0.0030						0.0016					
19.3		0.0040								0.0044											
20				0.1106	0.0008					0.1415			0.0308		0.0004	0.0906					0.0206

20.2									0.0004			0.0002			0.0012					
20.3		0.0006							0.0002											
21			0.0637						0.0911			0.0130		0.0002	0.1247					0.0006
21.2															0.0034					
22			0.0813						0.0741			0.0076			0.1785					0.0004
22.2															0.0046					
23			0.1306						0.0552			0.0028			0.1679					
23.2															0.0034					
23.3															0.0002					
24			0.1012						0.0180			0.0014			0.1619					
24.2															0.0048					
25			0.0685						0.0104			0.0004			0.1026					
25.2															0.0028					
25.3															0.0002					
26			0.0154						0.0014						0.0436					
26.2															0.0012					
26.3															0.0002					
27			0.0030												0.0137					
27.2															0.0002					
28			0.0010												0.0054					
29			0.0002												0.0026					
30															0.0002					
N	2504	2500	2504	2504	2504	2504	2494	2504	2500	2502	2485	2502	2497	2496	2504	2490	2235	2108	2502	2504	2499
Na	11	21	15	18	13	10	13	11	16	22	11	9	27	22	14	31	15	13	10	10	16
Ho	0.7492	0.8440	0.7380	0.8722	0.7496	0.7220	0.7927	0.8103	0.7536	0.8437	0.7666	0.7838	0.8614	0.8121	0.7364	0.8305	0.7808	0.4877	0.7450	0.6621	0.7943
He	0.7512	0.8893	0.7729	0.8902	0.7569	0.7567	0.8020	0.8305	0.7836	0.8665	0.8010	0.7983	0.8801	0.8227	0.7755	0.8728	0.8439	0.8803	0.7914	0.7048	0.8226
MP	0.1059	0.0224	0.0833	0.0226	0.1002	0.0978	0.0690	0.0499	0.0785	0.0324	0.0676	0.0697	0.0267	0.0509	0.0841	0.0287	0.0423	0.0420	0.0737	0.1341	0.0548
PE	0.5084	0.6831	0.4895	0.7391	0.5091	0.4632	0.5856	0.6183	0.5159	0.6825	0.5386	0.5693	0.7175	0.6217	0.4868	0.6569	0.5638	0.1769	0.5013	0.3722	0.5885
PD	0.8941	0.9776	0.9167	0.9774	0.8998	0.9022	0.9310	0.9501	0.9215	0.9676	0.9324	0.9303	0.9733	0.9491	0.9159	0.9713	0.9577	0.9580	0.9263	0.8659	0.9452
PIC	0.7104	0.8790	0.7393	0.8797	0.7169	0.7180	0.7726	0.8088	0.7501	0.8526	0.7738	0.7688	0.8680	0.8021	0.7421	0.8593	0.8244	0.8684	0.7589	0.6575	0.7985

N: number of samples; Na: number of alleles; Ho: Observed Heterozygosity; He: Expected Heterozygosity; MP: match probability; PE: power of exclusion; PD: power of discrimination; PIC: polymorphism information content.

Table 2 lists the allele frequencies and forensic parameters estimated for the whole dataset. The allele frequencies and forensic parameters estimated for each of the 26 populations (Supplementary Table S2) and the five population groups (Supplementary Table S3) are available as Supplementary Data. In general, the most polymorphic *loci* in all the populations were D1S1656, D2S1338, D12S391, D18S51, and FGA (Table 2). The analyzed *loci* were highly informative, with elevated PD values ranging between 86.59% (TPOX) and 97.76% (D1S1656). The combined MP was 5.72 × 10−27, and the combined PE was 0.99999997. Analysis of each locus in each population (Supplementary Table S2) showed that D22S1045 in PEL (71.61%) and D1S1656 in GBR (97.52%) presented the lowest and the highest PD value, respectively. The combined MP ranged from 1.98 × 10−25 in ACB to 2.20 × 10−21 in PEL.

We estimated the adherences of genotype frequencies to Hardy-Weinberg Equilibrium expectations for each STR marker at a population level (Table 3). Penta E presented heterozygote deficiency in 24 out of the 26 populations, leading to departures from the Hardy-Weinberg equilibrium. This finding indicated that HipSTR incorrectly called many heterozygous genotypes as homozygous. Disregarding Penta E, the number of deviations ranged from one (D13S317 and D16S539) to five (D19S433 and Penta D), and the number of deviations across populations ranged from zero (ASW and CEU) to seven (PUR), with an average of 2.42 departures in each population. When we considered the Bonferroni correction for multiple tests, only 39 departures remained significant, and most of them (61.53%) concerned PentaE.

Table 3. Probabilities of adherence to Hardy-Weinberg equilibrium proportions for each STR in all 26 subpopulations analyzed in the 1000 Genomes Project. Significant p-values (α = 0.05) are in boldface. The probabilities that remain significant after the Bonferroni correction for multiple tests (αBONFERRONI = 0.05/546 = 0.000092) are also underlined.

POP	CSF1PO	D1S1656	D2S441	D2S1338	D3S1358	D5S818	D7S820	D8S1179	D10S1248	D12S391	D13S317	D16S539	D18S51	D19S433	D22S1045	FGA	Penta D	Penta E	TH01	TPOX	vWA
ACB	0.8317	0.1003	0.9562	0.6768	0.7289	0.5439	**0.0238**	0.1626	0.7625	0.9937	0.1509	0.9058	0.1797	0.9795	0.8294	0.8814	0.1985	0.2895	0.4184	0.0678	0.1618
ASW	0.8494	0.8119	0.9805	0.7607	0.8298	0.8945	0.6251	0.9977	0.5225	0.4894	0.2447	0.9927	0.6351	0.3402	0.4035	0.8225	0.3519	0.0724	0.9511	0.9248	0.9409
BEB	0.6165	0.6321	0.1621	**0.0216**	0.9740	0.6614	0.4515	0.9476	0.8795	0.9124	0.5829	0.6662	0.3397	0.8489	0.6579	0.8494	0.4690	**0.0000**	0.2040	0.4850	0.9571
CDX	0.9917	0.3823	0.4668	0.9308	0.4053	**0.0000**	0.5757	0.9797	0.5425	0.4254	0.4678	0.2605	0.6701	0.5218	0.2454	0.1603	0.4560	**0.0000**	**0.0268**	**0.0025**	0.4988
CEU	0.9331	0.5301	0.9948	0.1233	0.1851	0.0674	0.6688	0.2600	0.8395	0.5314	0.8354	0.9559	0.1012	0.9471	0.8354	0.8997	0.5208	**0.0000**	0.7591	0.4191	0.1209
CHB	0.3105	0.7945	**0.0005**	0.8305	0.9407	0.8190	0.4159	0.2194	0.7585	**0.0003**	0.9664	0.4847	0.1581	**0.0047**	0.9689	0.9747	**0.0011**	**0.0000**	0.9424	0.8976	**0.0000**
CHS	0.2302	0.1077	0.2844	0.6237	0.5740	0.0727	0.9894	0.4883	0.3859	0.6180	0.7626	0.6386	0.3969	**0.0391**	0.8843	0.9618	0.4278	**0.0000**	0.1768	0.7723	0.8666
CLM	0.9075	0.3108	0.4415	0.0684	0.3560	**0.0470**	**0.0000**	0.8558	0.5450	**0.0004**	0.1412	0.5892	0.0829	0.9682	0.9999	0.9976	0.6915	**0.0000**	0.4287	0.4076	0.7427
ESN	0.1175	0.9988	0.9706	0.9750	**0.0303**	0.2028	0.6773	**0.0131**	0.8721	0.1069	0.8443	0.9823	0.9529	0.4869	**0.0209**	1.0000	**0.0110**	**0.0000**	0.9238	0.3436	0.0579
FIN	0.9558	0.4612	0.7627	**0.0000**	0.5922	0.5269	0.8596	0.3818	0.8976	0.9531	0.9688	0.7919	0.2147	0.9665	0.9869	0.7378	0.9930	**0.0000**	0.9504	0.8369	0.9465
GBR	0.9311	0.9506	0.2788	0.8505	0.9925	0.8037	0.8379	0.9828	0.8791	0.8061	0.2196	**0.0259**	0.2483	**0.0000**	**0.0317**	0.9879	0.2263	**0.0000**	0.0512	0.4530	0.4718
GIH	0.6965	**0.0239**	0.8370	0.6288	0.9993	0.4325	0.6899	**0.0011**	0.8836	0.9808	0.3863	0.6818	0.9979	0.7126	0.4995	0.7371	0.6790	**0.0000**	0.0770	0.7827	0.3344
GWD	0.1950	0.6832	0.9970	0.9978	0.5107	0.8942	0.2703	0.3213	0.9718	0.9527	0.4987	0.7810	**0.0098**	0.9973	0.2150	0.1932	**0.0000**	**0.0000**	0.1804	0.8530	0.9779
IBS	**0.0000**	0.8111	0.9373	0.9690	0.1246	0.9874	0.8882	0.8501	0.2448	0.5344	0.9012	0.1202	0.9943	0.9706	0.9373	0.5506	**0.0333**	**0.0000**	0.4393	0.6766	0.3111
ITU	0.6156	0.3367	**0.0097**	0.7459	0.7367	0.9723	0.8685	0.8721	0.1101	0.9617	0.7853	0.5636	0.9777	0.9016	0.9803	**0.0404**	0.8992	**0.0000**	0.9027	0.3481	0.4022
JPT	0.8554	0.7945	0.5003	0.7191	0.6566	0.6312	0.7590	0.2612	0.8154	0.7891	0.3862	0.9589	0.9922	0.9952	**0.0006**	0.7052	0.9159	**0.0000**	0.7771	**0.0002**	0.8666
KWV	0.9942	**0.0025**	0.8299	0.9795	0.9899	0.2980	0.2296	0.3737	0.7030	0.9483	0.5815	0.9489	0.1698	0.5107	0.9073	0.1862	**0.0000**	**0.0000**	0.5244	**0.0006**	0.4226
LWK	0.3621	0.9913	0.9838	0.3363	0.5245	0.8976	0.6751	0.6144	0.8436	0.1478	0.6934	0.7081	0.2947	0.4416	0.9396	0.9998	0.6011	**0.0000**	0.9548	**0.0325**	0.9572
MSL	0.8099	0.2244	0.4496	0.7258	0.9316	0.0628	**0.0372**	0.7088	0.0992	0.9442	0.9897	0.8470	1.0000	0.9398	0.2779	0.1057	0.2086	**0.0000**	**0.0171**	0.5908	**0.0000**
MXL	**0.0047**	**0.0109**	0.8297	0.6202	0.1254	0.8631	0.5456	**0.0125**	0.0509	0.4395	0.8989	0.7887	**0.0000**	0.3860	0.9233	0.1562	0.5281	**0.0000**	0.3747	0.4029	0.5094
PEL	0.8460	0.1247	0.9976	0.7122	0.1681	0.9781	0.7676	0.7192	0.9635	0.9402	0.1281	0.6186	0.9833	0.9176	0.8786	**0.0000**	0.5730	**0.0000**	0.8399	0.2170	**0.0467**
PJL	0.8683	**0.0073**	1.0000	0.6486	0.8966	**0.0001**	0.8996	0.5188	0.8722	0.5388	0.9943	0.4565	0.9976	0.6490	0.9166	**0.0053**	0.1485	**0.0000**	0.0721	0.6604	0.2985
PUR	0.7847	0.0819	**0.0058**	0.9097	0.1398	0.8342	0.7698	0.4704	**0.0034**	**0.0045**	**0.0337**	0.0556	0.7141	**0.0006**	0.9810	0.7965	0.7547	**0.0000**	**0.0000**	0.2571	**0.0191**
STU	0.9028	0.5930	**0.0000**	0.8290	0.2546	0.2661	**0.0071**	0.4770	0.9882	0.2049	0.0952	0.1927	0.4262	0.2335	**0.0001**	0.6382	0.2129	**0.0000**	0.6082	0.5311	0.9627
TSI	0.6299	0.4393	0.9777	0.0805	0.5719	0.4424	0.6240	0.4107	**0.0356**	0.6573	0.5777	0.9581	**0.0074**	**0.0000**	0.1240	0.5698	0.3556	**0.0000**	**0.0203**	0.3932	0.8569
YIR	0.6908	0.2395	0.8925	0.3004	**0.0000**	0.8611	0.7701	0.9277	0.6050	0.3079	0.5197	0.4061	0.9867	0.8417	0.9451	0.9643	0.4923	**0.0000**	0.4285	0.1299	0.9942

Principal Coordinates Analysis (PCoA) revealed four different population clusters (Figure 1). The first coordinate separated the cluster of African (AFR) populations on the right side. On the left side, we observed three different population groups: the European (EUR) populations in the upper part, the East Asian (EAS) populations in the lower section, and the South Asian (SAS) populations between them. The CLM, MXL, PEL, and PUR admixed populations clustered with the European populations, while the ACB and ASW populations clustered with the African (AFR) populations, reflecting their ancestry compositions.

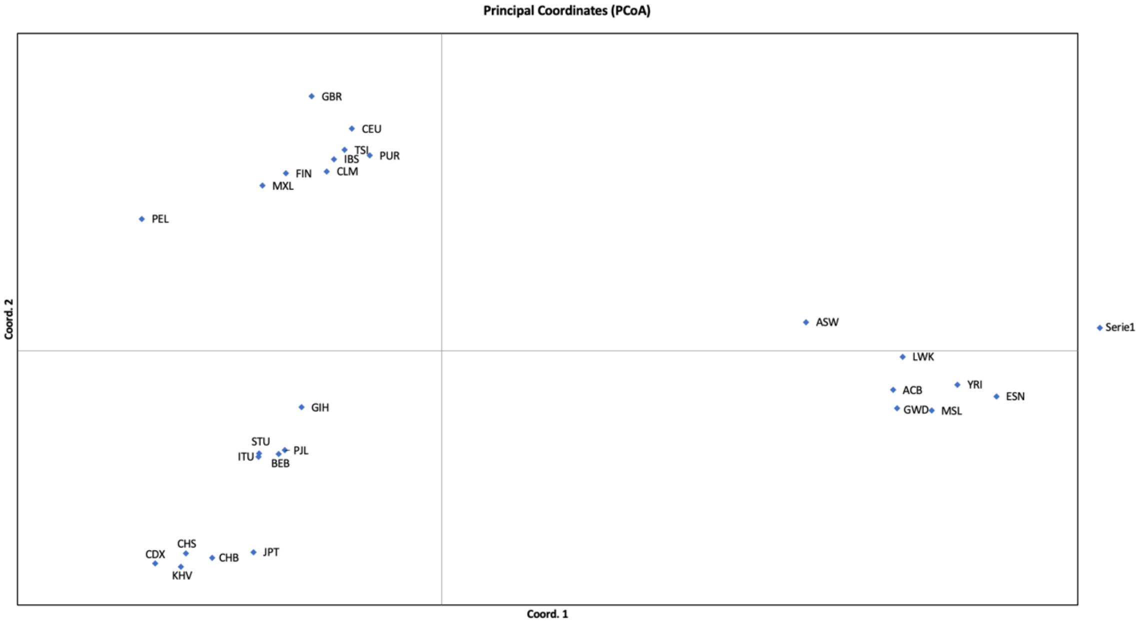

Figure 1. Principal Coordinates Analysis (PCoA) based on autosomal STR data regarding the 26 populations analyzed in the 1000 Genomes Project. Each point represents a population sample. More details on these populations are available in Supplementary Table S2. Coordinates 1 and 2 account for 39.15% and 19.25% of the variance, respectively.

We obtained similar results when we conducted the STRUCTURE analysis. Figure 2 depicts the STRUCTURE results derived from runs obtained with k ranging from three to six. When k = 4, each cluster reflected one of the major ancestries of the 1000 Genomes Project. Moreover, each of the six admixed American populations presented varying levels of ancestries from the four biogeographical groups. To verify the distribution of variance in different levels, we performed AMOVA by assuming a hierarchical structure that gathered the populations in four population groups: AFR, EAS, EUR, and SAS. We did not take the six populations in the AMR population group into account because their admixed compositions would bias the AMOVA results by reducing the proportion of variance between groups. We observed most of the variance within populations (97.12%). Differences between the four population groups accounted for 2.54% of the variance, whereas only 0.34% of the variance occurred due to differences between populations belonging to the same group.

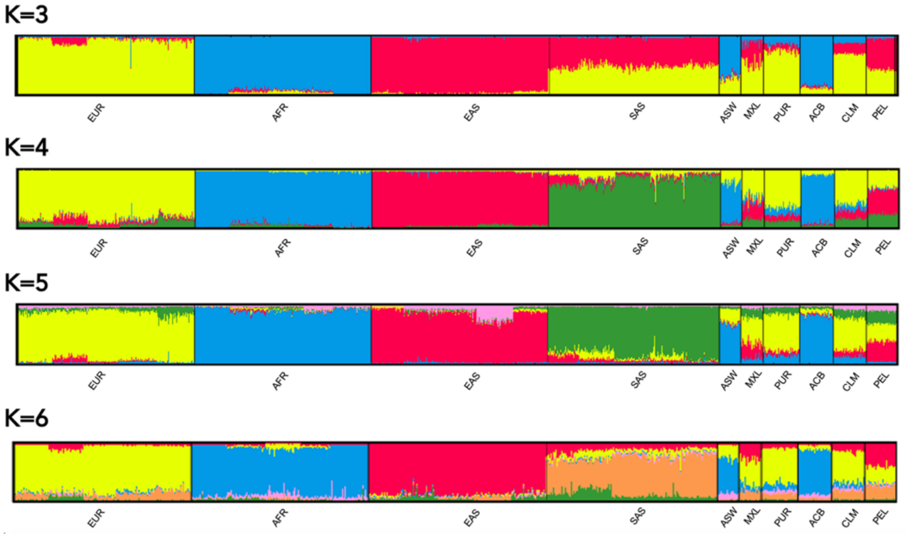

Figure 2. STRUCTURE analysis based on autosomal STR data obtained from the 26 populations included in the 1000 Genomes Project. Five sets of 100 independent runs, with the number of clusters ranging from 3 to 6, were conducted. Each bar plot depicts the results obtained from the run with the largest LnP (D) for the given k.

By using F_{ST}, we also compared the allele frequencies estimated from the dataset of the 1000 Genomes Project to the STR data retrieved for the same five major population groups (African, European, East Asian, South Asian, and admixed American) that composed the SPSmart STR browser (PopSTR) (Amigo et al., 2009) (Table 4). While the AMR (four), EAS (three), EUR (eight), and SAS (four) population groups presented small numbers of markers with significantly different frequencies between the two datasets, AFR presented 17 significant differences. This pattern might reflect the set of populations that compose the compared groups. Penta E was the only marker that showed significantly different F_{ST} values in all comparisons. By leaving AFR and Penta E aside, we observed only 15 significant differences out of 80 comparisons: the mean number of statistically significant differences was 0.75 per marker; this number ranged from zero (eight STR markers) to three (D2S441). When we considered the Bonferroni correction for multiple tests, only three of these 15 F_{ST} values remained significant, while six out of 16 significant differences observed for AFR (leaving Penta E aside), and all five Penta E differences remained significantly different.

Table 4. Probabilities obtained by F_{ST} of population differentiation comparing population groups from the 1000 Genomes Project with those from the SPSmart STR browser (Pop.STR) for each STR. Significant p-values (α = 0.05) are in boldface. The probabilities that remain significant after the Bonferroni correction for multiple tests ($\alpha_{BONFERRON}$I = 0.05/105 = 0.00048) are also underlined.

Marker	AFR		AMR		EAS		EUR		SAS	
	F_{ST}	*p*-value	F_{ST}	*p*-value	F_{ST}	*p*-value	F_{ST}	*p*-value	F_{ST}	*p*-value
CSF1PO	0.0084	**0.0057±0.0007**	-0.00104	0.7571±0.0047	-0.0014	0.6656±0.0053	0.0004	0.2445±0.0044	0.0019	0.1951±0.0036
D1S1656	0.0024	**0.0355±0.0019**	0.00028	0.3149±0.0046	-0.0024	0.9774±0.0015	0.0045	**0.0000±0.0000**	0.0021	0.1152±0.0032
D2S441	0.0093	**0.0005±0.0002**	0.00370	**0.0216±0.0015**	0.0025	0.1251±0.0027	0.0074	**0.0000±0.0000**	0.0122	**0.0058±0.0007**
D2S1338	0.0124	**0.0000±0.0000**	-0.00048	0.6527±0.0049	-0.0026	0.9940±0.0007	0.0031	**0.0008±0.0003**	0.0024	0.1044±0.0030
D3S1358	0.0044	**0.021±0.0014**	-0.00122	0.8369±0.0040	0.0010	0.2578±0.0044	0.0013	0.0835±0.0026	0.0033	0.1143±0.0031
D5S818	0.0030	**0.0477±0.0024**	0.00609	**0.0051±0.0008**	-0.0014	0.6936±0.0048	-0.0007	0.7638±0.0047	-0.0014	0.6250±0.0050
D7S820	0.0047	**0.0095±0.0009**	-0.00105	0.7982±0.0041	-0.0006	0.5151±0.0050	-0.0003	0.5677±0.0054	-0.0023	0.8731±0.0034
D8S1179	0.0033	**0.0308±0.0018**	-0.00158	0.9817±0.0014	-0.0016	0.7986±0.0042	0.0024	**0.0150±0.0012**	-0.0003	0.4615±0.0049
D10S1248	0.0017	0.1081±0.0033	-0.00132	0.8984±0.0029	-0.0031	0.9976±0.0005	0.0003	0.2627±0.0044	-0.0001	0.4055±0.0046
D12S391	0.0007	0.1920±0.0041	0.00077	0.1805±0.0036	-0.0019	0.8774±0.0030	0.0008	0.0946±0.0030	-0.0009	0.6155±0.0047
D13S317	0.0160	**0.0000±0.0000**	0.00017	0.3478±0.0044	-0.0018	0.7993±0.0045	0.0003	0.2904±0.0047	0.0051	**0.0354±0.0019**
D16S539	0.0105	**0.0002±0.0001**	-0.00051	0.5773±0.0047	-0.0010	0.5939±0.0046	0.0047	**0.0009±0.0003**	-0.0010	0.5987±0.0052
D18S51	0.0018	**0.0483±0.0019**	-0.00138	0.9831±0.0013	-0.0007	0.5795±0.0055	0.0005	0.2028±0.0042	-0.0003	0.4671±0.0047
D19S433	0.0093	**0.0003±0.0002**	0.00151	0.1085±0.0027	0.0010	0.2476±0.0041	0.0000	0.3759±0.0049	0.0036	0.0660±0.0028
D22S1045	0.0004	0.2943±0.0053	0.00217	0.0766±0.0025	0.0094	**0.0055±0.0007**	0.0034	**0.0087±0.0009**	0.0054	0.0516±0.0021
FGA	0.0011	0.1503±0.0037	-0.00089	0.8382±0.0034	0.0036	**0.0397±0.0020**	0.0007	0.1306±0.0035	0.0018	0.1518±0.0039
Penta D	0.0275	**0.0000±0.0000**	0.00318	**0.0113±0.0010**	-0.0015	0.7660±0.0042	0.0003	0.2969±0.0042	0.0000	0.4117±0.0048
Penta E	0.0139	**0.0000±0.0000**	0.00733	**0.0000±0.0000**	0.0412	**0.0000±0.0000**	0.0083	**0.0000±0.0000**	0.0179	**0.0000±0.0000**
TH01	0.0103	**0.0004±0.0002**	0.00090	0.2083±0.0036	-0.0025	0.9262±0.0027	0.0057	**0.0001±0.0001**	0.0026	0.1425±0.0038
TPOX	0.0039	**0.0269±0.0019**	-0.00057	0.5622±0.005	-0.0011	0.5341±0.0046	0.0015	0.0939±0.0033	0.0063	**0.0455±0.0020**
vWA	0.0065	**0.0007±0.0003**	-0.00095	0.7575±0.0047	0.0005	0.3135±0.0047	-0.0001	0.4312±0.0051	-0.0010	0.5977±0.0049

4. DISCUSSION

The present study provides the most diverse database of forensic autosomal STR markers obtained from global populations. STR markers display high levels of polymorphism, which makes them attractive for forensic purposes and population genetics studies. This is the first time that the 1000 Genomes high-coverage (~30x) dataset has been used for STR genotyping purpose. Although a few previous initiatives (Gymrek et al., 2012; Tang et al., 2017; Willems et al., 2014) attempted to genotype forensically relevant STRs, they only dealt with previous low-coverage 1000 Genomes releases (~7.4x), which prevented the acquisition of results or resulted in highly unreliable genotypes due to large rates of allele dropout. Moreover, it should be emphasized that even the last paper that presented the high-coverage WGS data did not include STR variants in the results and stated that genotyping STRs from such data remains a considerable challenge Byrska-Bishop (Byrska-Bishop et al., 2022).

In forensic genetics, STR markers consist in the most widespread and informative tool for human identification. In spite of the limitations addressed below, such as unreliability of Penta D and Penta E genotypes involving specific alleles, this NGS-based STR database presents reliable allele frequencies that could be used in criminal casework to estimate the rarity of a given STR-based profile from a query sample of unknown or uncertain ancestry in various worldwide populations. This could instantly, and without additional costs, trigger a DNA-based intelligence strategy to guide enquiries (West et al., 2020) providing hints and/or assigning biogeographical origin in many situations, such as a missing person investigation (West et al., 2020; Pereira et al., 2011), leaving only the most complex cases for supplementary analysis with a most suitable set of Ancestry Informative Markers.

Short-read next generation sequencing is slowly being introduced in forensic labs worldwide. Although such technology is still restricted and expensive, it has become more sensitive, requiring as little as 25 pg of extracted DNA, and is suitable to solve more complex cases, such as discrimination of twins (using STRs, WGS or mtDNA sequencing approaches) and deconvolution of highly unbalanced mixtures reviewed by Carratto et al., (2022). Some criminal (Yuan et al.,2020; Diepenbroek et al., 2020; Knijf et al., 2020), kinship (Pilli et al., 2022) and missing persons (Cuenca et al., 2020) casework already benefiting from this have been reported. However, genotyping STR markers by using NGS data, especially WGS assays, may be challenging—accurate genotyping requires high coverage, longer alleles are difficult to detect due to reads of limited sizes, and mutations in flanking regions may lead to null alleles (Aalbers et al., 2020). These and other issues have been addressed by Gaag et al. (2016) and Valle-Silva et al. (2022).

Notwithstanding the challenges addressed here, several studies have demonstrated that STRs can be genotyped by using dedicated bioinformatics tools. Software such as LobSTR (Gymrek et al., 2012), toaSTR (Ganschow et al., 2018), STRait Razor (Warshauer et al., 2013), and HipSTR (Willems et al., 2017), among others, have shown consistent and accurate results (Valle-Silva et al., 2022; Halman et al., 2020). Moreover, Bornman et al. (2012) demonstrated that, by using an NGS approach, CODIS *loci* could be accurately called even from mixtures.

Particularly for the deconvolution of mixtures, the identification of isometric alleles (i.e., alleles with the same length but containing different repeat sequences) is a necessary task, since it further increases the discriminating power of the currently used STR markers; nevertheless, it is not achieved with traditional PCR and capillary electrophoresis techniques (Alvarez-Cubero et al., 2017; Ballard et al., 2020). This sequence-based analysis is already feasible with small-scale targeted sequencing assays, particularly those using kits and software solutions tailored for forensic purposes, such as the ForenSeq DNA Signature Prep Kit coupled with the ForenSeq™ Universal Analysis Software (Verogen Inc., San Diego, CA, USA) or the Precision ID GlobalFiler™ NGS STR Panel v2 coupled with the Converge Software NGS Analysis Module (Thermo Fisher Scientific), but it is still a challenge for large-scale WGS assays. In order to achieve this goal concerning big data in the near future, new bioinformatics tools must be developed, or the current ones further improved.

Willems et al. (2014) analyzed human STR variation by using lobSTR. These authors employed the data of Phase 1 of the 1000 Genomes Project. The data were generated by using low-sequencing coverage, which is excessively error-prone. In fact, the authors reported difficulties in detecting both alleles in each sample, which resulted in an overall deficit of heterozygotes. As previously addressed, several reasons led us to choose HipSTR to call STR genotypes from this high-coverage dataset of the 1000 Genomes Project. Because HipSTR allows the flanking regions to be customized, almost any STR marker can be evaluated in hundreds of samples at once. At first glance, HipSTR may appear more complex, but it is the most appropriate tool to deal with whole genomes. In addition, a recent evaluation of the performance of this tool revealed high efficiency and accuracy levels (Valle-Silva et al., 2022).

Although HipSTR provides flexibility, the major limitation of this study is the inability to genotype D21S11, which is one of the 20 CODIS *loci*. Additional limitations are the failure in detecting two very small Penta D alleles and the biased allele frequencies of very large Penta D and Penta E alleles probably because of sequence-specific features, such as the GC content (Wang et al., 2011; Sims et al., 2014; Castelli et al., 2017) producing low depth of coverage bias and/or the limited length of the Illumina NGS reads (150 bp paired-end reads). This issue could be immediately circumvented with long-read sequencing technologies, such as those implemented in Pacific Biosciences (PacBio) and Oxford Nanopore platforms. However, one should not expect that long-read sequencing would be suitable for a wide range of forensic samples, which are often degraded and/or available in low amounts (Wang et al., 2011; Sims et al., 2014; Castelli et al., 2017; Belsare et al., 2022). It is noteworthy that, by employing 300 nucleotide-long paired-end reads in a targeted sequencing assay, we successfully genotyped D21S11 with HipSTR, which suggests a sequencing methodology issue rather than a bioinformatics issue (Valle-Silva et al., 2022).

In this study, Penta D and Penta E showed 10.74% and 15.81% of missing data, respectively. By using Illumina sequencing technology, van der Gaag et al. (2016) showed that longer alleles of Penta D, Penta E, and FGA presented sequencing errors at the end of the reads, which resulted in null alleles and genotyping errors. As observed for D21S11, this issue was probably related to the impossibility of detecting longer alleles due to read-length constraints. Furthermore, we did not detect two very small Penta D alleles (2.2 and 3.2), which are common in African populations, which was unexpected. Supplementary Table S4 compares the allele frequencies estimated in the present study with the allele frequencies obtained from the SPSmart STR browser (PopSTR) (Amigo et al., 2009) for the major population groups. Such straightforward comparison showed that we were not able to detect alleles larger than 18 in Penta E. This failure led directly to Hardy-Weinberg equilibrium deviations (Table S3) due to deficit of heterozygotes in 24 out of the 26 studied populations. Thus, allele frequencies estimated for Penta E were strongly biased toward increased frequencies of shorter alleles and have limited applicability (Supplementary Table S2). The probabilities obtained with the F_{ST} analysis (Table 4) supported this conclusion: Penta E presented significant F_{ST} values in all five comparisons. Although Penta D and FGA also posed this problem, their undetected alleles usually have low frequencies—Except for Penta D alleles 2.2 and 3.2 in African populations (Supplementary Table S4). Therefore, this technical issue did not influence the Hardy-Weinberg equilibrium and F_{ST} analysis as much as Penta E. Although this comparison is valid and helpful, we must emphasize that the compared samples corresponded to distinct population groups. The African population group in popSTR comprised mainly East African Somalian individuals (404 out of 507 samples), while the African populations in the 1000 Genomes Project samples corresponded to West Africa. Similarly, over 50% of the European population group in popSTR was composed mainly of U.S. Europeans (1443 out of 2135) (Clarke et al., 2017; Amigo et al., 2009). Taken together, these results attest that the bioinformatics analysis performed in the present study is robust, and that the distribution of allele frequencies is reliable for all *loci* except Penta E.

The most polymorphic *loci* in the whole dataset of the 1000 Genomes Project were D1S1656, D2S1338, D12S391, D18S51, and FGA. All these markers presented high degrees of polymorphism throughout the world. AMOVA revealed that most of the variance (97.12%) in allele frequencies occurred within populations, corroborating previous studies (Rosenberg, 2011; Rosenberg et al., 2002). A study that evaluated human population structure using genotypes at 377 autosomal microsatellite *loci* in 1056 individuals from 52 worldwide populations revealed that the variance within populations accounts for 93 to 95% of genetic variation, while differences among major groups constitute only 3 to 5% (Rosenberg et al., 2002; Jobling, 2022). Although the number of populations and genetic markers are quite different, the larger amount of variance within populations and lower variance among groups observed in the present study may be either due to chance or to the fact that forensic STRs do show relatively lower F_{ST} than random STRs due to the increased heterozygosity of the former (Jobling, 2022). However, as expected, AMOVA, together with principal component analysis (Figure 1) and the clustering analysis performed with STRUCTURE (Figure 2), confirmed that the four ancestral populations groups (AFR, EUR, EAS, and SAS) defined by the 1000 Genomes Consortium did differ significantly from each other. Given that the admixed American populations present different ancestry compositions (Figure 2), most of them clustered with Europeans, while ACB and ASW clustered with Africans (Figure 1).

The results obtained with the STRUCTURE software corroborated the relationship between the different population groups and provided additional support for the reliability of the calculated genotypes. When k = 3, SAS resembled an admixture between EAS and EUR. A specific cluster for SAS emerged when k = 4. When k = 5, a minor Eurasian (shared between EUR and EAS) component arose. When k = 6, the SAS-shared ancestry with EUR and EAS became more evident. Regarding the admixed American populations, irrespective of the number of clusters considered, ACB and ASW revealed their preeminent African origin, CLM and PUR revealed more extensive European ancestry, and MXL and PEL revealed almost equal amounts of European and Amerindian (i.e., EAS) ancestries. These results fully corroborated the distribution of the populations into the PCoA (Figure 1). Additional clusters did not provide increased resolution with straightforward meaning.

The outcome of this population genetics evaluation further corroborates the robustness and reliability of this STR dataset. Despite all the applications already addressed in the beginning of this section, the most important contribution of this open access genotype dataset probably lies in the fact that it may be used to estimate and establish additional population genetics parameters that may be taken as direct references in many studies that are using the 1000 Genomes Project dataset to retrieve new sets of SNPs, indels and microhaplotypes in various efforts to maximize intelligence from DNA evidence (de la Puente et al., 2021; Phillips et al., 2020; Lan et al., 2020; Huang et al., 2022).

5. CONCLUSIONS

We were able to offer a reliable open-access STR database based on the high-coverage (30x) WGS data of Phase 3 of the 1000 Genomes Project generated by the NYGC. However, the limited length of sequencing reads introduces noticeable bias in allele frequencies estimated for Penta D and Penta E. The reliability of this dataset is supported by (a) previous studies attesting that HipSTR is efficient, (b) the Hardy-Weinberg equilibrium analysis, (c) the set of analyses employed to evaluate the interpopulation genetic diversity, and (d) the comparison between the allele frequencies obtained here and the frequencies obtained by other initiatives that used capillary electrophoresis. Although we expect that this openaccess database will be of great interest for future forensic studies on population genetics, the current 1000 Genomes Project dataset does not describe human genetic diversity worldwide. In fact, many biogeographical regions, mainly in Oceania and the Americas, have not been sampled, indicating that additional large-scale initiatives may provide further insight into STR diversity in populations worldwide.

6. SUPPLEMENTARY MATERIALS

The following supporting information can be downloaded at: https: //www.mdpi.com/article/10.3390/genes13122205/s1.

7. AUTHOR CONTRIBUTIONS

T.S.F. and G.V.-S. contributed equally to this work; conceptualization, investigation, methodology, analysis, and writing of the paper. J.A. supported and helped with bioinformatic tools (software), and C.T.M.-J. with writing—reviewing, editing, and supervising. All authors have read and agreed to the published version of the manuscript.

8. FUNDING

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior-Brasil (CAPES)-Finance Code 001. C.T.M.-J. (#312802/2018-8) is supported by a Research fellowship from CNPq/Brazil.

9. INSTITUTIONAL REVIEW BOARD STATEMENT

Ethical review and approval were waived for this study because all data were derived from the 1000 genomes public database.

10. INFORMED CONSENT STATEMENT

Not applicable.

11. DATA AVAILABILITY STATEMENT

1000 Genomes Project Phase 3 samples in a high-coverage (30x) assay using the NovaSeq 6000 Sequencing System (Illumina, Inc.) https://www.internationalgenome.org/data-portal/data-collection/30x-grch38 accessed on 10 July 2021.

12. ACKNOWLEDGMENTS

We thank Thomas Willems for sharing his knowledge and expertise in bioinformatics and population genetics and his support with the HipSTR tool, and Cynthia Maria de Campos Prado Manso for her assistance with the English language.

13. CONFLICTS OF INTEREST

The authors declare no conflict of interest.

14. REFERENCES

Aalbers, S.E.; Hipp, M.J.; Kennedy, S.R.; Weir, B.S (2020). Analyzing population structure for forensic STR markers in next generation sequencing data. Forensic Sci. Int. Genet. 49,102364. Doi: 10.1016/j.fsigen.2020.102364

Alvarez-Cubero, M.J.; Saiz, M.; Martínez-García, B.; Sayalero, S.M., et al (2017). Next generation sequencing: An application in forensic sciences? Ann. Hum. Biol. 44, 581–592. Doi: 10.1080/03014460.2017.1375155.

Amigo J.; Christopher, P.; Toño, S.; Fernandez, F.L., et al (2009). pop.STR—An online population frequency browser for established and new forensic STRs. Forensic Sci. Int. Genet. Suppl. Ser. 2, 361–362. Doi: 10.1016/j.fsigss.2009.08.178

Ballard, D.; Winkler-Galicki, J.; Wesoły, J (2020). Massive parallel sequencing in forensics: Advantages, issues, technicalities, and prospects. Int. J. Leg. Med. 134, 1291–1303. Doi:10.1007/s00414-020-02294-0

Belsare, S.; Levy-Sakin, M.; Mostovoy, Y.; Durinck, S., et al (2019). Evaluating the quality of the 1000 genomes project data. BMC Genom. 20, 620. Doi: 10.1186/s12864-019-5957-x.

Bornman, D.M.; Hester, M.E.; Schuetter, J.M.; Kasoji, M.D., et al (2012). Short-read, high-throughput sequencing technology for STR genotyping. Biotech. Rapid Dispatches 1–6. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4301848/

Børsting, C.; Morling, N (2015). Next generation sequencing and its applications in forensic genetics. Forensic Sci. Int. Genet. 18, 78–89. Doi: 10.1016/j.fsigen.2015.02.002

Byrska-Bishop, M.; Evani, U.S.; Zhao, X.; Basile, A.O., et al (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell 2022, 185, 3426–3440.e3419. Doi: 10.1016/j.cell.2022.08.004

Carratto, T.M.T.; Moraes, V.M.S.; Recalde, T.S.F.; Oliveira, M.L.G., et al (2022). Applications of massively parallel sequencing in forensic genetics. Genet. Mol. Biol. 45, e20220077. Doi: 10.1590/1678-4685-GMB-2022-0077

Castelli, E.C.; Gerasimou, P.; Paz, M.A.; Ramalho, J., et al (2017). HLA-G variability and haplotypes detected by massively parallel sequencing procedures in the geographicaly distinct population samples of Brazil and Cyprus. Mol. Immunol. 83, 115–126. Doi: 10.1016/j.molimm.2017.01.020

Clarke, L.; Fairley, S.; Zheng-Bradley, X.; Streeter, I., et al (2017). The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data. Nucleic Acids Res. 2017, 45, D854–D859. Doi: 10.1093/nar/gkw829

Cuenca, D.; Battaglia, J.; Halsing, M.; Sheehan, S (2020). Mitochondrial Sequencing of Missing Persons DNA Casework by Implementing Thermo Fisher's Precision ID mtDNA Whole Genome Assay. Genes 11(11):1303. Doi: 10.3390/genes11111303

de la Puente, M.; Ruiz-Ramírez, J.; Ambroa-Conde, A.; Xavier, C., et al (2021). Development and Evaluation of the Ancestry Informative Marker Panel of the VISAGE Basic Tool. Genes 12, 1284. Doi: 10.3390/genes12081284

Diepenbroek, M.; Bayer, B.; Schwender, K.; Schiller, R., et al (2020). Evaluation of the Ion AmpliSeq™ PhenoTrivium Panel: MPS-Based Assay for Ancestry and Phenotype Predictions Challenged by Casework Samples. Genes 2020, 11, 1398. Doi: 10.3390/genes11121398

Excoffier, L.; Lischer, H.E (2010). Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Mol. Ecol. Resour. 10, 564–567. Doi: 10.1111/j.1755-0998.2010.02847.x

Fairley, S.; Lowy-Gallego, E.; Perry, E.; Flicek, P (2020). The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. Nucleic Acids Res. 48, D941–D947. Doi: 10.1093/nar/gkz836

Fungtammasan, A.; Ananda, G.; Hile, S.E.; Su, M.S., et al (2015). Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. Genome Res. 25, 736–749. Doi: 10.1101/gr.185892.114.

Ganschow, S.; Silvery, J.; Kalinowski, J.; Tiemann, C (2018). toaSTR: A web application for forensic STR genotyping by massively parallel sequencing. Forensic Sci. Int. Genet. 37, 21–28. Doi: 10.1016/j.fsigen.2018.07.006.

Gettings, K.B.; Ballard, D.; Bodner, M.; Borsuk, L.A., et al (2019). Report from the STRAND Working Group on the 2019 STR sequence nomenclature meeting. Forensic Sci. Int. Genet. 43, 102165. Doi: 10.1016/j.fsigen.2019.102165.

Gouy, A.; Zieger, M (2017). STRAF-A convenient online tool for STR data evaluation in forensic genetics. Forensic Sci. Int. Genet. 30, 148–151. Doi: 10.1016/j.fsigen.2017.07.007.

Gymrek, M.; Golan, D.; Rosset, S.; Erlich, Y (2012). lobSTR: A short tandem repeat profiler for personal genomes. Genome Res. 22, 1154–1162. Doi: 10.1101/gr.135780.111.

Halman, A.; Oshlack, A (2020). Accuracy of short tandem repeats genotyping tools in whole exome sequencing data. F1000Res 9, 200. Doi: 10.12688/f1000research.22639.1

Huang, S.; Sheng, M.; Li, Z.; Li, K., et al (2022). Inferring bio-geographical ancestry with 35 microhaplotypes. Forensic Sci. Int. 341, 111509. Doi: 10.1016/j.forsciint.2022.111509

Hubisz, M.J.; Falush, D.; Stephens, M.; Pritchard, J.K (2009). Inferring weak population structure with the assistance of sample group information. Mol. Ecol. Resour 9(5), 1322–1332. Doi: 10.1111/j.1755-0998.2009.02591.x

Jobling, M.A (2022). Forensic genetics through the lens of Lewontin: Population structure, ancestry and race. Philos. Trans. R. Soc. Lond. B Biol. Sci. 2022, 377, 20200422. Doi: 10.1098/rstb.2020.0422

Knijf, P.D. How Next Generation Sequencing Resolved a Difficult Case, Leading to the First Criminal Conviction of Its Kind; Verogen: San Diego, CA, USA, 2020; pp. 1–4. Available at: https://verogen.com/wp-content/uploads/2020/12/ngs-first-criminal-conviction-case-study-vd2019024-b.pdf

Lan, Q.; Fang, Y.; Mei, S.; Xie, T., et al (2020). Next generation sequencing of a set of ancestry-informative SNPs: Ancestry assignment of three continental populations and estimating ancestry composition for Mongolians. Mol. Genet. Genom. 295, 1027–1038. Doi: 10.1007/s00438-020-01660-2

Peakall, R.; Smouse, P.E (2012). GenAlEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research-an update. Bioinformatics 28, 2537–2539. Doi: 10.1093/bioinformatics/bts460

Pereira, L.; Alshamali, F.; Andreassen, R.; Ballard, R., et al (2011). PopAffiliator: Online calculator for individual affiliation to a major population group based on 17 autosomal short tandem repeat genotype profile. Int. J. Leg. Med. 125, 629–636. Doi: 10.1007/s00414-010-0472-2.

Phillips, C.; Amigo, J.; Tillmar, A.O.; Peck, M.A., et al (2020). A compilation of tri-allelic SNPs from 1000 Genomes and use of the most polymorphic loci for a large-scale human identification panel. Forensic Sci. Int. Genet. 46, 102232. Doi: 10.1016/j.fsigen.2020.102232.

Pilli, E.; Tarallo, R.; Riccia, P.; Berti, A., et al (2022). Kinship assignment with the ForenSeq™ DNA Signature Prep Kit: Sources of error in simulated and real cases. Sci. Justice 62, 1–9. Doi: 10.1016/j.scijus.2021.10.007

Robinson, J.T.; Thorvaldsdóttir, H.; Wenger, A.M.; Zehir, A., et al (2017). Variant Review with the Integrative Genomics Viewer. Cancer Res. 77, e31–e34. Doi: 10.1158/0008-5472.CAN-17-0337

Rosenberg, N.A (2011). A population-genetic perspective on the similarities and differences among worldwide human populations. Hum. Biol. 83, 659–684. Doi: 10.3378/027.083.0601.

Rosenberg, N.A (2004). Distruct: A program for the graphical display of population structure. Mol. Ecol. Notes 4, 137–138. Doi: 10.1046/j.1471-8286.2003.00566.x.

Rosenberg, N.A.; Pritchard, J.K.; Weber, J.L.; Cann, H.M., et al (2002). Genetic structure of human populations. Science 298, 2381–2385. Doi: 10.1126/science.1078311.

Sims, D.; Sudbery, I.; Ilott, N.E.; Heger, A., et al (2014). Sequencing depth and coverage: Key considerations in genomic analyses. Nat. Rev. Genet. 15, 121–132. Doi: 10.1038/nrg3642

Sudmant, P.H.; Rausch, T.; Gardner, E.J.; Handsaker, R.E., et al (2015). An integrated map of structural variation in 2,504 human genomes. Nature 526, 75–81. Doi: 10.1038/nature15394.

Tang, H.; Kirkness, E.F.; Lippert, C.; Biggs, W.H., et al (2017). Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes. Am. J. Hum. Genet. 101, 700–715. Doi: 10.1016/j.ajhg.2017.09.013

Thermo Fisher Scientific. Precision ID GlobalFiler™ NGS STR Panel v2. Available online: http://www.thermofisher.com/hid-ngs (accessed on 20 October 2022).

Thorvaldsdóttir, H.; Robinson, J.T.; Mesirov, J.P (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief. Bioinform. 14, 178–192. Doi: 10.1093/bib/bbs017.

Valle-Silva, G.; Frontanilla, T.S.; Ayala, J.; Donadi, E.A., et al (2022). Analysis and comparison of the STR genotypes called with HipSTR, STRait Razor and toaSTR by using next generation sequencing data in a Brazilian population sample. Forensic Sci. Int. Genet. 58, 102676. Doi: 10.1016/j.fsigen.2022.102676

van der Gaag, K.J.; de Leeuw, R.H.; Hoogenboom, J.; Patel, J., et al (2016). Massively parallel sequencing of short tandem repeats-Population data and mixture analysis results for the PowerSeq™ system. Forensic Sci. Int. Genet. 24, 86–96. Doi: 10.1016/j.fsigen.2016.05.016

Verogen. Universal Analysis Software. Available online: https://verogen.com/products/universal-analysis-software/ (accessed on 20 October 2022).

Wang, W.; Wei, Z.; Lam, T.W.; Wang, J (2011). Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions. Sci. Rep. 1, 55. Doi: 10.1038/srep00055

Warshauer, D.H.; Lin, D.; Hari, K.; Jain, R., et al (2013). STRait Razor: A length-based forensic STR allele-calling tool for use with second generation sequencing data. Forensic Sci. Int. Genet. 7, 409–417. Doi: 10.1016/j.fsigen.2013.04.005

West, F.L.; Algee-Hewitt, B.F.B (2020). Cadaveric blood cards: Assessing DNA quality and quantity and the utility of STRs for the individual estimation of trihybrid ancestry and admixture proportions. Forensic Sci. Int. Synerg. 2, 114–122. Doi: 10.1016/j.fsisyn.2020.03.002.

Willems, T.; Gymrek, M.; Highnam, G.; Mittelman, D., et al (2014). The landscape of human STR variation. Genome Res. 24, 1894–1904. Doi: 10.1101/gr.177774.114.

Willems, T.; Zielinski, D.; Yuan, J.; Gordon, A., et al (2017). Genome-wide profiling of heritable and de novo STR variations. Nat. Methods 14, 590–592. Doi: 10.1038/nmeth.4267

Yuan, L.; Chen, X.; Liu, Z.; Liu, Q., et al (2020). Identification of the perpetrator among identical twins using next-generation sequencing technology: A case report. Forensic Sci. Int. Genet. 44, 102167. Doi: 10.1016/j.fsigen.2019.102167.

1000 Genomes Project Consortium, Auton A, Brooks L.D, Durbin R.M., et al (2015). A global reference for human genetic variation. Nature 1;526(7571):68-74. Doi: 10.1038/nature15393.

CONCLUSÃO

O conjunto de marcadores STR aqui utilizados mostrou ser efetivo para estimar a estrutura populacional e a ancestralidade em níveis populacional e individual. Apesar da menor diversidade interpopulacional característica destes marcadores, os resultados obtidos se mostraram perfeitamente alinhados ao conhecimento pré-existente relacionado à história demográfica das populações estudadas.

As validações realizadas utilizando genótipos determinados por eletroforese capilar, bem como bases de dados de frequências alélicas dos mesmos grupos populacionais (PopSTR) revelaram alta acurácia e concordância nos três artigos apresentados.

O HipSTR se mostrou ser uma ferramenta adequada para genotipar marcadores STR a partir de dados de NGS e a mais indicada para lidar com grandes conjuntos de dados genômicos. Entretanto, o uso de mais de um software contribui para aumentar a acurácia principalmente nos marcadores que HipSTR não mostrou ser muito efetivo (D21S11, Penta D, Penta E).

Mais estudos serão realizados com esta ferramenta para tentar melhorar a captura de alelos nos diferentes marcadores, buscando inclusive identificar isoalelos a partir de dados de NGS, o que consiste em uma grande vantagem dessa técnica sobre metodologias tradicionais.

MATERIAIS SUPLEMENTARES

Capítulo 1

Analysis and comparison of the STR genotypes called with HipSTR, STRait Razor and toaSTR by using next generation sequencing data in a Brazilian population sample

Materiais suplementares disponíveis:
https://www.sciencedirect.com/science/article/pii/S1872497322000175#sec0065

Capítulo 2

Materiais suplementares disponíveis:
https://drive.google.com/drive/folders/1NDtyDsTbC1IvLM_OlMFZO5ZO1PIkrD3T?usp=share_link

Capítulo 3

Open-Access Worldwide Population STR Database Constructed Using High-Coverage Massively Parallel Sequencing Data Obtained from the 1000 Genomes Project

Materiais suplementares disponíveis:
https://www.mdpi.com/article/10.3390/genes13122205/s1